HOCHSCHULE
MITTWEIDA
UNIVERSITY OF
APPLIED SCIENCES

MASTER THESIS

Mr
Florian Heinke

Energy profile-based function analysis of globular and membrane proteins

2012

Department of **Mathematics, Natural and Computer sciences**

MASTER THESIS

Energy profile-based function analysis of globular and membrane proteins

Author:
Florian Heinke

Course of studies:
Molecular Biology/Bioinformatics

Seminar group:
MO10w1-M

First examiner:
Prof. Dr. rer. nat. Dirk Labudde

Second examiner:
M.Sc. Steffen Grunert

Mittweida, August 2012

Let the truth of love be lighted
Let the love of truth shine clear
Sensibility
Armed with sense and liberty
With the heart and mind united
In a single
Perfect
Sphere

Cygnus X-1, Book II: Hemispheres
Rush

Bibliographic description

Heinke, Florian: Energy profile-based function analysis of globular and membrane proteins, 111 pages, 3 figures, Hochschule Mittweida (FH), Department of Mathematics, Natural and Computer sciences

Master Thesis, 2012

Abstract

Proteins are macromolecules that consist of linear-bonded amino acids. They are essential elements in various metabolic processes. The three-dimensional structure of a protein is determined by the order of amino acids, also referred to as the protein sequence. This conformation corresponds to the structural state in which the protein is functionally active. However, relationships between protein sequence, structure and function have not been fully understood yet. Additionally, information about structural properties or even the entire protein structure are crucial for understanding the dynamics that define protein functionality and mechanisms. From this, the role of a protein in its molecular context can be described closely. For instance, interactions can be investigated and comprehended as a biological dynamic network that is sensitive to alternations, i.e. changes which are caused by diseases. Such knowledge can aid in drug design, whereas compounds need to be specifically tailored and adjusted to their molecular targets.
Protein energy profile-based methods can be applied to investigate protein structures concerning dynamics and alternations. The publications enclosed to this work discuss in general the scientific potentials of energy profile-based techniques and algorithms. On the one hand, changes in stability caused by protein mutations and protein-ligand interactions are discussed in the context of energy profiles. On the other hand, energetic relations to protein sequence, structure and function are elucidated in detail. Finally, the presented discussions focus on recent enhancements of the eProS (energy profile suite) database and toolbox. eProS freely provides all elucidated methodologies to the scientific community. Thus, one can address biological questions with the presented methods at hand. Additionally, eProS provides annotations related to foreign databases. This ensures a broad view on biological data and information. In particular, energetic characteristics can be identified which contribute to a protein's structure and function.

Zusammenfassung

Proteine sind Makromoleküle, die sich aus linear verknüpften Aminosäuren zusammensetzen. Sie bilden essenzielle Elemente bei der Realisierung metabolischer Prozesse. Die dreidimensionale Struktur entspricht der funktionell aktiven Konformation eines Proteins, wobei diese eindeutig durch die Sequenz - die Aminosäureabfolge - definiert ist. Jedoch sind die Zusammenhänge zwischen Proteinsequenz, Struktur und Funktion noch nicht eindeutig verstanden. Zudem sind Informationen bezüglich struktureller Eigenschaften oder der kompletten Struktur eines Proteins essenziell für das Verständnis der Dynamiken, die Funktionalität und Mechanik definieren. Daraus lässt sich wiederum die Rolle des Proteins im molekularen Kontext erschließen - welche Interaktionen ein Protein eingeht und wie das Netzwerk molekularer Interaktionen bedingt durch z.B. krankhaften Veränderungen entartet und ggf. medikamentös beeinflusst werden könnte.
Proteinenergieprofile stellen eine Möglichkeit dar, Proteinstrukturen bezüglich Dynamik und Alternationen hin zu untersuchen. Die mit dieser Arbeit vorliegenden wissenschaftlichen Veröffentlichungen beleuchten im wesentlichen das Potential energieprofilbasierter Analyse-Techniken und Algorithmen. Zum einen werden Stabilitätsveränderungen, die durch Proteinmutationen und Protein-Ligand-Interkationen bedingt werden, auf Grundlage von Energieprofilen diskutiert. Zum andern werden Zusammenhänge zwischen Energie, Sequenz, Struktur und Funktion näher erläutert. Abschließend wird auf Arbeiten an der eProS (energy profile suite) Datenbank und Toolbox eingegangen. eProS bietet Wissenschaftlern die Möglichkeit sämtliche, in dieser Arbeit erläuterten Methoden auf ihre Fragestellungen anzuwenden. Weiterhin bietet eProS Annotationen aus etlichen Fremdquellen an, wodurch eine Vielzahl von Informationen bereitgestellt wird. Diese können die Identifizierung von charakteristischen, energetischen Merkmalen ermöglichen, die die Ausprägung des untersuchten Charakteristikums (z.B. strukturelle oder funktionelle Eigenschaften) bedingen.

Acknowledgments

First of all, I am especially grateful to Professor Dirk Labudde at the University of Applied Sciences Mittweida for giving me the opportunity to work on this subject. His way of approaching, analyzing and digging to the hearts of biological matters always helped me keeping my mind straight. Very special thanks go to Daniel Stockmann. Without the numerous hours of joyful discussions and his eye-opening advices, I would probably still be programming in a black box. I would also like to take the opportunity to thank Steffen Grunert and Michael Spranger for their very helpful hints and support in programming and TEX-ing. Many thanks go to all the people who loved to share their knowledge but, sadly, cannot be named here explicitly. You know who you are.
Finally, I would like to thank my family for their encouragements, never-ending patience and support.

I. Contents

II. List of Figures

1 Introduction and Subject Description

The development and improvement of technologies applied in laboratory experiments lead to an even faster growing amount of biological data. Over the decades, bioinformatics has become the key in analyzing, processing and storing data [1, 2]. For instance, answering the seemingly trivial question of "Is the nucleic acid sequence, I obtained from this genome, already present in a database?" is biologically, computationally and mathematically ambitious. On the one hand, thousands or even millions of sequences need to be processed - a procedure that stresses computational feasibility. On the other hand, the run time needs to be handleable in practice and algorithms have to be as sensitive as possible. Thus it is not surprising that BLAST [3], the most applied program package for these purposes, has undergone twenty years of improvements to fulfill the increasing day-to-day requirements [4].

However, biological processes can only be understood if data is contemplated as the outcome of interactions and dynamics which transpire at molecular or cellular level. Here, as essential elements, proteins realize a variety of functions, e.g. mediating molecular signals, ensuring nutrient and ion flow, catalyzing chemical reactions and regulating gene expression [5]. The experimental determination of a protein's sequence as well as sequence analyses are well established methods. However, to determine functional features and protein dynamics, the employment of demanding in vivo/in vitro and in silico techniques is required. Essentially, protein function and activity are determined by the protein's structure. Thus, once structural data have been generated from spectroscopy experiments (e.g. circular dichroism [6], nucleic magnetic resonance [7] or X-ray [2, 7]), functional features can be explored in detail [1]. But experimental methods of structure determination of proteins (especially membrane proteins and large complexes) are limited, since the necessary processes of protein isolation, enrichment and crystallization are resource-demanding and challenging [8]. These cutbacks lead to discrepancies in information which are reflected by the number of entries found in public databases. More precisely, compared to the non-redundant protein sequence dataset stored at the National Center of Biotechnology Information (NCBI, ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz) [9] which currently holds almost 20 million sequences, the Protein Data Bank (PDB) [10] only stores about 11,000 non-redundant protein chain structures (computed by PDB-REPRDB [11]) - a number almost evanescent in comparison to the amount of available sequence data. Since the correspondence of sequence to structure and function are not fully understood, closing this gap of information is one of the biggest goals in modern biology. Additionally, determining a protein's dynamics and process-regulatory influences in molecular interaction networks are tasks that are even more complex than the seemingly trivial question of querying a sequence database.

For solving the protein structure determination problem, computational approaches have been developed which all rely on force fields and energy models for approximating atom-

atom or residue-residue interaction potentials. Over the years, numerous models have been developed, ranging from complex all-atom interaction models based on physics laws to coarse-grained energy models. In coarse-grained energy models, physical information are smoothed, for instance by abstracting single molecules or monomers (e.g. amino acid residues or nucleotides) into single spatial points, by simplifying potential calculation by employing methods of statistical physics, by using straight-forward interaction schemes or by discretizing continuous spaces into lattice-like spatial systems. Physics-based energy models are mostly applied in molecular dynamics simulations (MD) or ab initio protein folding. However, MD simulations, which take all-atom information into account, are computationally demanding. Thus, detailed molecular simulations are limited to relatively small systems, e.g. proteins up to a size of 150 residues. Herein lies the strongest advantage of coarse-grained energy models, since information smoothing goes along with a reduction of system complexity and, hence, a reduction of computational demands [12, 13].

In 2006 Kozielski and colleagues proposed that sequences of residue energies computed from protein structures utilizing potential functions can be applied to identify proteins with common protein family membership. Furthermore, the group discussed that protein-environment interactions can be investigated. As their basic methodology, Kozielski and colleagues introduced modified Needleman-Wunsch [14] and Smith-Waterman [15] algorithms for aligning energy profiles. However, Kozielski et al based their studies on physics-based approaches by utilizing the TINKER [16] MD software suite [17, 18, 19]. Although TINKER is a relatively straight-forward software compared to other suites (i.e. NAMD2 [20]) for computing residue-wise potentials, it is still quite error-prone, slow, only semi-automatable, very sensitive to structural abnormalities (such as chain breaks due to low experimental X-ray resolution) and non-canonical interactions. Thus, high-throughput analyses and generating large-scale databases of physics-based energy profiles is challenging. In 2007 F. Dressel implemented a straight-forward coarse-grained energy model that relies on basic spatial residue distributions and concepts of statistical physics [21, 22]. This energy model has been applied for protein folding studies as well as structure-based investigations of data derived from single-molecule force spectroscopy (SMFS).

I got introduced to this model in late 2009 as part of my course work, whereas concepts of a so-called super alignment were discussed. Besides including protein sequence and secondary structure information, information of residue energies derived by the aforementioned coarse-grained model had to be considered as well. The super alignment had been proposed and discussed as an alternative for deriving protein structure similarities by including these sources of information. Additionally, due to its familiarity to classic sequence alignment algorithms, the super alignment had shown a computational complexity that had been superior to the established C_α atom-based, physico-chemical insensitive structure alignment procedures (i.e. Dalilite [23] or FATCAT [24]). As part of my bachelors thesis, energy profiles had been investigated concerning several properties and correlations to structural features. For instance correspondences between energy profile progression and secondary structure elements had been investigated.

Furthermore, based on the techniques employed in the super alignment algorithm a method for aligning energy profiles had been developed and discussed. It had been shown that, similar to Kozielski and colleagues, coarse-grained energy profiles can be applied and aligned to identify similar protein structures. In addition, this energy profile alignment procedure (eAlign) had shown equal performance and sensitivity compared to the super alignment algorithm. Based on this it had been proposed that energy profiles can be interpreted as protein-specific abstractions of physico-chemical and structural information as well as fingerprints that are descriptive for a protein's function and structure. As a further result, a modified GOR algorithm had been implemented and improved for performing energy profile prediction from sequence. In addition a scoring scheme had been proposed that reflects protein energy profile similarity as a measure of pseudo-distance. In the studies presented in this thesis, this scheme is referred to as dScore. However, at this point correlations of similar energy profiles derived by eAlign to similar protein function had to be affirmed manually. Furthermore, the energy model had to be investigated concerning correlations to energy values derived by physics-based approaches to verify its interpretability as a model for describing protein/residue energy and stability.

For these investigations 220 non-redundant globular protein structures (pairwise sequence identity $<$25%) have been retrieved from the PDB. For each structure the physics-based (fine-grained) total potential energy (E_{FG}) has been computed by utilizing TINKER. Subsequently, coarse-grained energy profiles have been computed and all energy values of each profile have been add to obtain the total coarse-grained energy E_{CG}. The plot of the observed correlation between both methods for energy computation is shown in Figure 1.1. The Pearson correlation coefficient has been found to be $\rho = 0.95$. Thus, the coarse-grained energy model approximates physics-based energy potentials very well and, hence, the model can be applied for the analysis of protein dynamics, stabilizing aspects influencing or accounting protein function as well as studying effects caused by protein-environment interactions.

To prove correlations of dScore relations to sequence-structure-function relationships on a large-scale basis, multiple processing steps have been made. First, a set of 2,700 non-redundant protein structures has been obtained from the PDB. Employing the ClustalX standalone software package [25] has led to all pairwise sequence identities. All pairwise structure alignments have been performed by utilizing the heuristic MAMMOTH algorithm [26] implementation available by the MaxCluster software [27]. Because of the large number of resulting alignments and imbalance of non-significant to significant alignments, Monte Carlo-sampling has been employed with the product of sequence identity and TMscore computed by MaxCluster serving as acceptance constraint. This sampling led to 80,000 alignments with balanced-distributed biological significance. In the process, the resulting alignments have been re-calculated by utilizing a self-written implementation of the Needleman-Wunsch algorithm and the rigidFATCAT structure alignment algorithm. For computing functional similarity the G-SESAME [28] web server has been employed. Basically, G-SESAME locates each entry of two Gene Ontology (GO) [29] term sets, that have been annotated to a protein or biological entity in

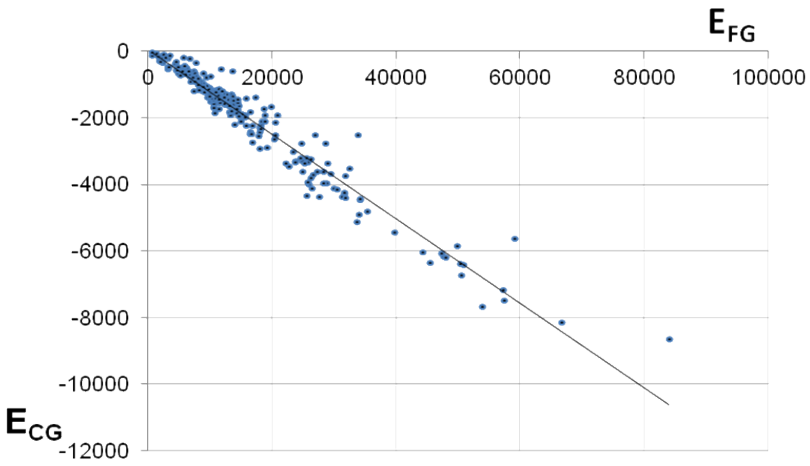


Figure 1.1: 220 non-redundant globular protein structures had been sampled from the Protein Data Bank [10] and, subsequently, total potential energies have been computed by a fine-grained, physics-based model (corresponding to E_{FG}) and the coarse-grained energy model discussed in this work (corresponding to E_{CG}). As shown, energy values correlate well ($\rho = 0.95$). Hence, the coarse-grained energy model is applicable for biological and physical meaningful discussions and interpretations, f.e. investigating destabilizing effects caused by point mutations or protein-environment interactions.

general, in the Gene Ontology. By abstracting the Gene Ontology as an acyclic directed graph and utilizing graph theory algorithms, common paths of terms to the ontology root can be identified and weighted. From this, a scoring is derived, that, as shown by Du et al [28], corresponds to functional similarity. For each alignment the corresponding energy profile alignment and dScore have been finally computed. As shown in Figure 1.2, dScores correlate to sequence, structure and functional similarity.

The peer-reviewed papers and bock chapters, conference presentations and conference posters as well as publications submitted for peer-review listed in this thesis were written between August 2010 and August 2012 and reflect the continuous work on this subject.

As a methodology for investigating sequence-structure-relationships, functional divergences and similarities, energetic properties which determine protein function and protein family memberships as well effects of external forces causing (de)stabilizations in protein structure and dynamics, the analysis of protein energy profiles can contribute in understanding and predicting the molecular properties of a protein which define its role in biological networks. Newly gained knowledge can aid and contribute in comprehending the molecular mechanism that control and trigger biological processes which have not been understood yet. In August 2010, the eProS database and toolbox has been established as a repository for energy profile data and sandbox for analysing, predicting, aligning and searching protein energy profiles. eProS is freely available for the scientific community and can contribute to their work by providing various techniques of working

with energy profiles.

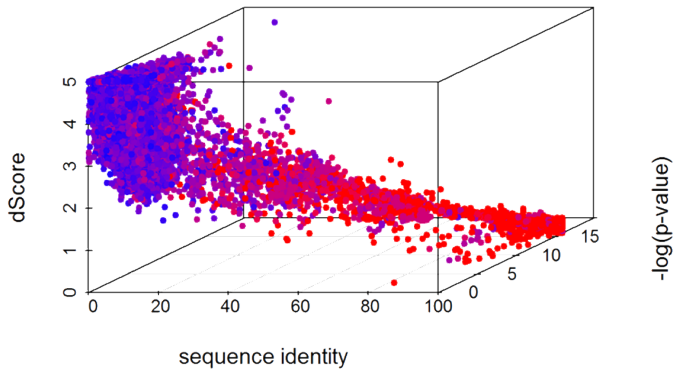


Figure 1.2: To investigate correlations of energy profile similarity (reflected by a pseudo-distance referred to as dScore), non-redundant protein structures had been sampled from the PDB and sequence identities as well as structure similarities have been computed. Additionally, functional similarity of proteins has been determined by means of the G-SESAME web server [28] which determines functional similarity semantically using Gene Ontology [29] terms. In this plot, results from pairwise sequence, structure and energy profile alignments as well as semantic functional similarity analyses are depicted. Structure alignment scores are given as p-values at $-log$-scale. Semantic similarity is illustrated by blue-to-red coloring. Blue-colored alignments correspond to pairs of protein structures with no common function. This plot demonstrates that energy profiles yield sequence and structure information. Additionally, energy profile similarities reflected by dScores correlate to functional similarities.

2 Publications

2.1 Book Chapter and Peer-reviewed Paper

2.1.1 Analysis of Membrane Protein Stability in Diabetes insipidus

The book chapter presented in this section was written in April 2011. It has been edited by Kyuzi Kamoi and published by Intech electronically with the ISBN number 978-953-307-367-5 and DOI number 10.5772/22258 electronically and as hard copy. After its release in November 2011 it achieved a cumulative number of about 750 downloads until August 2012. The pie chart depicted in Figure 2.1 illustrates the accumulative download share of the countries from which this chapter has been downloaded the most. The chapter deals with the analysis membrane proteins that are essential in water reabsorption through the apical membrane of kidney cells. Corresponding dysfunctional mutant proteins have been identified as causes for a rare disorder known as nephrogenic diabetes insipidus. In the chapter, the target proteins, aquaporin-2 and V2R, are discussed in detail and mutants are investigated by means of energy profile-based methods. A multiple energy profile alignment algorithm (MEPAL) is employed in the analysis. Theoretical results are in agreement with experimental findings. Thus, the theoretical approaches have substantiated the proposed mechanisms in the investigated proteins.

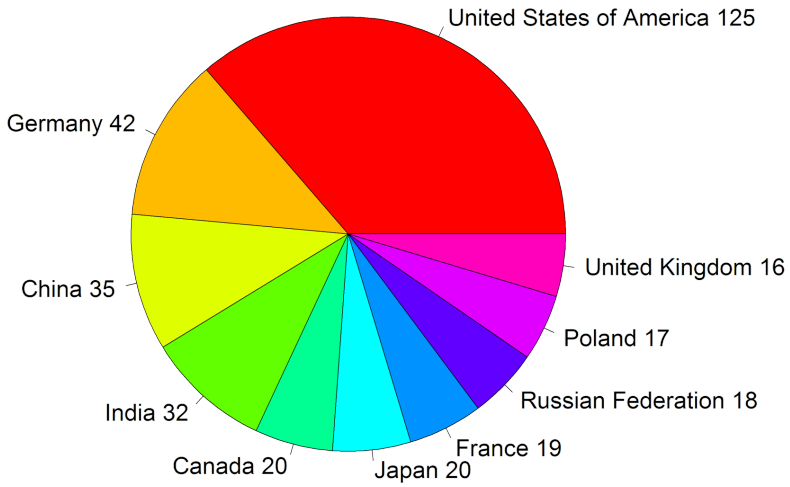


Figure 2.1: Download share of the book chapter by country. The countries represented by this pie chart are the ten countries from which this chapter has been accessed the most.

5

Analysis of Membrane Protein Stability in *Diabetes Insipidus*

Florian Heinke, Anne Tuukkanen and Dirk Labudde
University of Applied Sciences Mittweida
Germany

1. Introduction

Diabetes insipidus (DI) is a rare endocrine disorder, with an incidence in the general population assessed on one case per 25,000-30,000 people (Robertson, 1995; Ananthakrishnan, 2009; Krysiak, et al., 2010). It is a disease characterized by polyuria and compensatory polydipsia. The underlying causes of DI are diverse and can be central **defects**, in which no functional arginine-vasopressin is released from the pituitary, or may becaused by defects in the kidney (nephrogenic DI, NDI). Four different types of NDI are known. First, acquired NDI can originate as a side-effect of drugs, with the most prominent being the antibipolar drug lithium. Second and third, autosomal recessive and dominant inheritable NDI, are caused by gene mutations in the AQP2 gene encoding aquaporin-2. Finally, mutations in the AVPR2 gene (Deen et al., 1994; Mulders, 1998), which encodes the V2 vasopressin receptor (V2R), are the cause of the X-linked inheritable form of NDI (Fig. 1 right) (Van den Ouweland et al., 1992; Rosenthal, 1992).

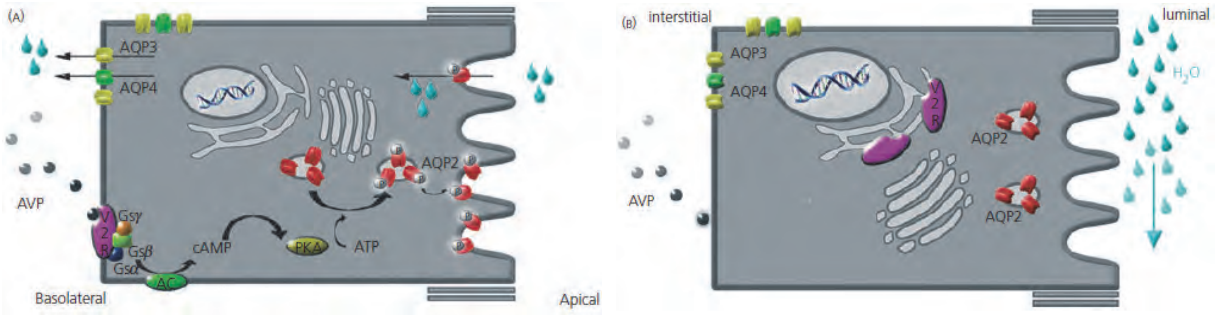


Fig. 1. Transcellular water transport in renal collecting principal duct cells and molecular cause of X-linked nephrogenic diabetes insipidus (NDI)

(Left) Vasopressin binding to its type 2 receptor (V2R) triggers a cAMP cascade that leads to the insertion of aquaporin-2 water channels in the apical membrane. This allows water to pass through this membrane and transcellular water transport to balance concentration of the pro-urine and, there by antidiuresis. (Right) In the X-linked form, NDI is often caused by V2R mutants trapped in the endoplasmic reticulum as a result of their misfolding, making them unavailable for binding arginine-vasopressin (AVP) at the basolateral plasma membrane. As a result, no transcellular water transport takes place, leading to polyuria (Los et al., 2010).

The X-linked inheritable form of nephrogenic diabetes insipidus is a disorder in which patients are unable to concentrate their urine despite the presence of the hormone arginine-vasopressin (AVP). This antidiuretic hormone regulates the process of the water reabsorption, according to the body's need, from the pro-urine that is formed by ultrafiltration in the kidney. It binds to its type 2 receptor in the kidney (Fig. 1, A). Mutations in the gene encoding the V2R often lead to NDI. Many of these mutations do not interfere with the intrinsic functionality of V2R, but cause its retention in the endoplasmic reticulum (ER) making it unavailable for AVP binding.
As a consequence of the inability of the kidneys to concentrate the pro-urine in response to AVP, diseased adult patients may have a daily output of 15–20 l of highly dilute (usually < 100 mOsmol / kg) urine. In newborn infants, NDI is characterized by irritability, poor feeding, poor weight gain and dehydration symptoms.
Classically, the diagnosis NDI was made after a dehydration test (Los et al., 2010) but it has become possible in clinical practice to apply direct analysis of the arginine vasopressin V2 receptor gene (AVPR2) and the aquaporin-2 gene for the diagnosis and differential diagnosis of nephrogenic diabetes insipidus (Fujiwara & Bichet, 2005).
To date, over 200 mutations have been described in the AVPR2 gene, which can be categorized into classes according to their cellular fate (Robben et al. 2006).
Another gene for the diagnosis of **Diabetes insipidus** is WFS1. It encodes a transmembrane protein which induces the Wolfram Syndrom (Hardy et al. 1999), a rare autosomal recessive disorder characterized by juvenile-onset non-autoimmune Diabetes mellitus, optic atrophy, sensorineural deafness and **Diabetes insipidus** (Wolfram & Wagener, 1938). In addition, psychiatric illnesses such as depression and impulsive behavior are frequently observed in affected individuals (Swift & Swift, 2001).
The minimal criteria for diagnosis are Diabetes mellitus and optic atrophy. **Diabetes insipidus**, sensorineural deafness, urinary tract anatomy, ataxia, peripheral neuropathy, mental retardation and psychiatric illness are additional symptoms seen in the majority of patients (Strom, 1998a).
The WFS1 protein, also called wolframin, consists of 890 amino acids and was predicted to have nine or ten membrane spanning domains (Inoue et al., 1998; Strom et al., 1998b). More than 100 mutations of the WFS1 gene have been identified to date in Wolfram syndrome patients. Most are inactivating mutations, suggesting loss of function to be responsible for the disease phenotype (Cryns et al., 2003). The WFS1 protein is expressed in various tissues but at higher levels in the brain, heart, lung and pancreas (Inoue et al., 1998; Strom et al., 1998b). The literature shows that the WFS1 protein is to be localized predominantly in the endoplasmic reticulum and suggested a possible role of this protein in membrane trafficking, protein processing and/or regulation of cellular calcium homeostasis (Takeda et al, 2001). A recent study showed this protein to contain nine transmembrane domains and to be embedded in the ER membrane with the amino-terminus in the cytosol and the carboxy-terminus in the ER lumen (Hofmann et al., 2003).
The short introduction shows the correlation of **Diabetes insipidus** with mutations in different membrane proteins. Membrane proteins play essential roles in cellular processes. Despite the central importance of transmembrane proteins, the number of high-resolution structures remains small due to the practical difficulties in crystalizing them. Many human disease-linked point mutations occur in transmembrane proteins. These mutations cause structural instabilities in a transmembrane protein leading it to unfold or missfold in an alternative conformation.

However, the analysis of this stability plays an important part concerning the understanding process of these diseases, especially for **Diabetes insipidus.** In this chapter, we demonstrate two different approaches on membrane protein stability analysis, results from single-molecule force spectroscopy (SMFS) on aquaporin-1 and a new method based on so called energy profiles.

2. Description of the investigated membrane proteins

The points of interest in this work are the membrane proteins: aquaporins -2, -3 and -4 as well as the arginine vasopressin V2 receptor.

2.1 Aquaporins

For a better understanding of the relationship between mutations and changes in the stability of membrane proteins, we summarize in this section the structural characteristics of water cannels. Knowledge of these aquaporins derived by experimental data revealed the affiliation of aquaporins to a family of related water channels from many species. Aquaporins provide highly permeable pores for water to cross membranes. Four identical subunits form a stable tetramer in the plane of membrane. Each subunit has a narrow pore that is selective for water passing through the middle of a bundle of α-helices. About 10 water molecules line up in a pore about 0.3 nm in diameter. Hydrophobic bonding of water with a pair of asparagine residues (Fig. 2: Asn 76 and Asn 192, human aquaporin-1 numbering) at a narrow point in the pore allows the channel to be selective for water. The monomers of the protein arose by gene duplication, since their sequences are remarkably similar. Various human tissues express 12 different aquaporin isoforms. Aquaporin-1 (Fig. 2) is found in red blood cells, retinal proximal tubules, blood vessel endothelial cells, and the choroid plexus. Aquaporin-2 is required for renal collecting ducts to reabsorb water (King et al. 2004).

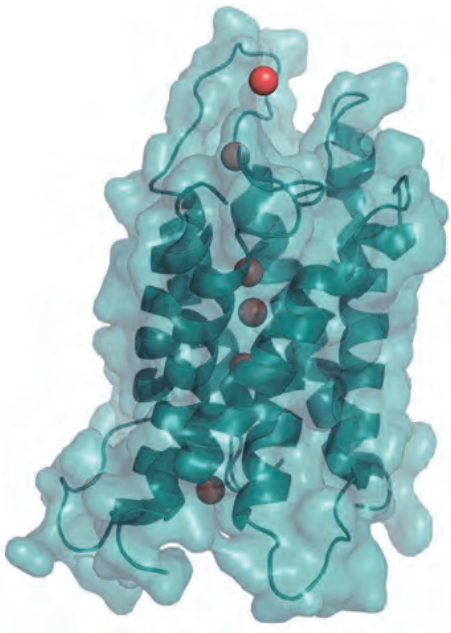


Fig. 2. Structure of an aquaprin-1 protein (PDB_ID 1ih5); The lined up water molecules are shown in red. The protein structure contains seven α-helices.

Antidiuretic hormone controls the insertion of aquaporin-2 in the collecting duct membrane. It activates a seven-helix receptor, causing cytoplasmic vesicles storing aquaporin-2 to fuse with the plasma membrane. This increases the permeability of apical plasma membranes to water, allowing it to move from the urine into the hypertonic extracellular space of the renal medulla. The water selectivity can appreciate by the protein structure. The figure 3 illustrates the network of the involved residues in the process. Additional to the exposed residue pair (Asn 76 and Asn 192) we observed Arg 195 and His 180 on the top of the pore, both closing the pore for bigger molecules or ions. The pore is hydrophobic inside. The peptide bonds of the residues Gly 188 and Ile 191 can form h-bonds with water molecules. The reaction of sensitive Cys 189 with mercuric ions closes the water pore (King et al. 2004; Pollard & Earnshaw, 2007).

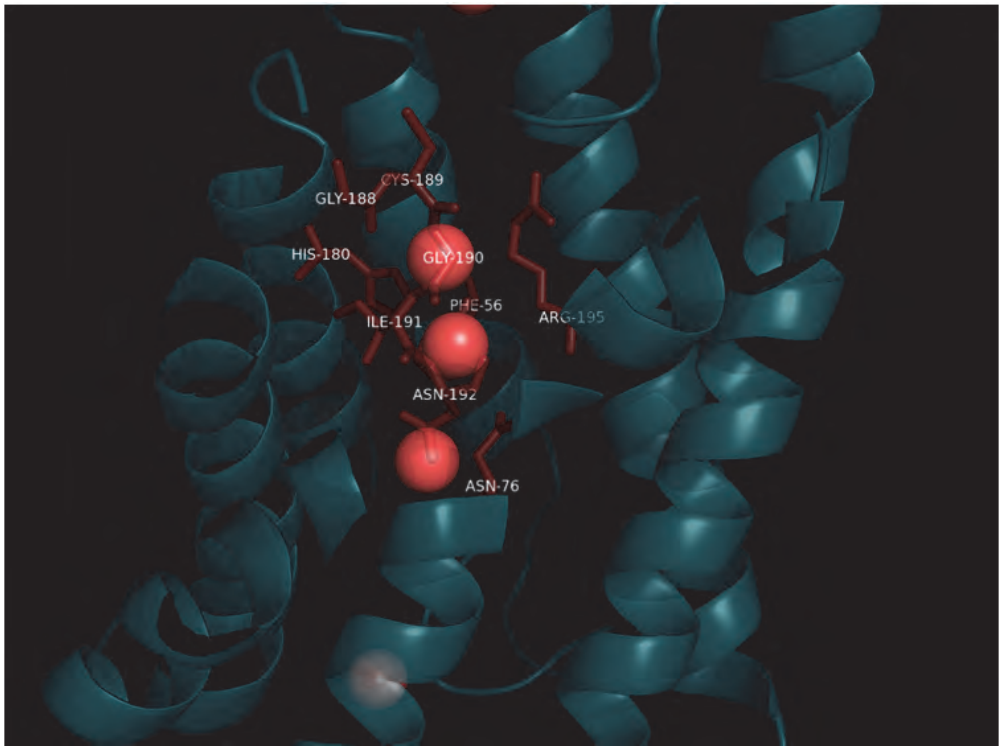


Fig. 3. Aquaporin 1 and the residue network for water transport. The following residues are involved: His 180, Gly 188, Cys 189, Gly 190, Ile 191, Asn 192, Arg 195 and Phe 56, Asn 76.

Furthermore, it is shown that two of the seven helices hold a highly conserved Asn-Pro-Ala motif (Chen et al., 2006). These two motifs meet in opposite α-helical orientation. This conformation induces a bipolar electric field changing the water molecule orientation and preventing protons to move through the channel. Further molecular simulation study has revealed a secondary free energy-barrier induced by Phe 56, His 180 and Arg 195. This barrier is located at the extracellular side, about 8 Å apart from the bipolar field. It forms a constriction region with a diameter of approximately 2 Å which allows only a single water molecule to pass the pore. Thus, the secondary free-energy barrier plays a main role in transport selectivity. Additionally, molecular dynamics simulation of Arg 195 mutants showed a significant decrease of the secondary energy-barrier leading to the loss of selectivity. This indicates that conformational changes or mutations of Arg 195 have a main influence on the transport behavior of aquaporin (de Groot et al. 2004; Chakrabarti et al., 2004a; Chakrabarti et al., 2004b; Ilan et. al, 2004).

The focus of this chapter is the analyses of the stability of membrane proteins. To collect a reliable dataset and to gather information about existing protein structures, we checked the Protein Data Bank (PDB) and the ModBase for aquaporin entries. Table 1 and table 2 give an overview over the used structures for all future calculations and discussions of the aquaporins.

Aquaporin (PDB_ID)	Sequence length (PDB)	Sequence length (Uniprot)	Number missing residues N-terminal	Number missing residues C-terminal	Coverage [%]
Aqp1/1fqy	230	273	7	36	84
Aqp4/3gd8	227	328	31	70	69
Aqp5/3d9s	251	269	1	17	93

Table 1. Overview of aquaporin structure

Aquaporin	BLAST-hit in PDB	e-Value	Model	Model-Template	reliability	Sequence identity
Aqp2:	3d9sA	2.00E-99	x	3d9sA	good	68%
Aqp3:	**1ldfA**	2.00E-46	x	3ldfA	average	43%

Table 2. Overview of models in the ModBase for the structural unknown aquaporin proteins. The highlighted (bold) PDB_ID 1ldf is the structure of a glycerol channel from *E.coli*.

While the structures of aquaporin-1, -4 and -5 have been clarified by electron crystallography (aquaporin-1) or x-ray diffraction (aquaporin-4 and -5) respectively, the structures of aquaporin-2 and -3 have been predicted by homology modeling. In homology modeling, the sequence of a structural unknown protein is queried to a protein structure database (such as the Protein Data Bank or Protein Data Bank of Transmembrane Proteins). The structure with adequate sequence identity (usually greater than 25-30%, depending on the length of the query sequence) is used as the modeling template. By simulations with force fields, using rotamer libraries and machine learning techniques, the query sequence is modeled into the given structure template and the resulting model can be evaluated. Unsuccessful modeling is caused by low sequence identity and leads to short modeled fragments or no model at all. Many of those successfully modeled structures are stored and organized in protein model databases (*e.g.* Modbase and Protein Model Portal). Because of the relatively low number of known structures finding an appropriate template is still a bottleneck in structure homology modeling.

As seen in table 2, the structure model of aquaporin-2 has been produced by using the structure of aquaporin-5 as modeling template. The most reliable structure of aquaporin-3 was modeled on the basis of a glycerol channel of *E. coli.* The neighbor joining tree (see figure 4) of aquaporin 1-5 and the glycerol channel gives insight to sequence similarity of the involved proteins and, in case of aquaporin-2 and -3 their modeling template sequences. As seen by branch length, aquaporin-2 and aquaporin-5 share the highest sequence identity in the entire tree which confirms the applicability of the aquaporin-5 structure of modeling a high reliable aquaporin-2 structure. Aquaporin-3 and the glycerol channel form a single isolated monophyletic cluster with a branch length of about 600 indicating the moderate sequence identity of 43% given in table 2. However, this sequence identity is high enough for deriving a model with an average reliability.

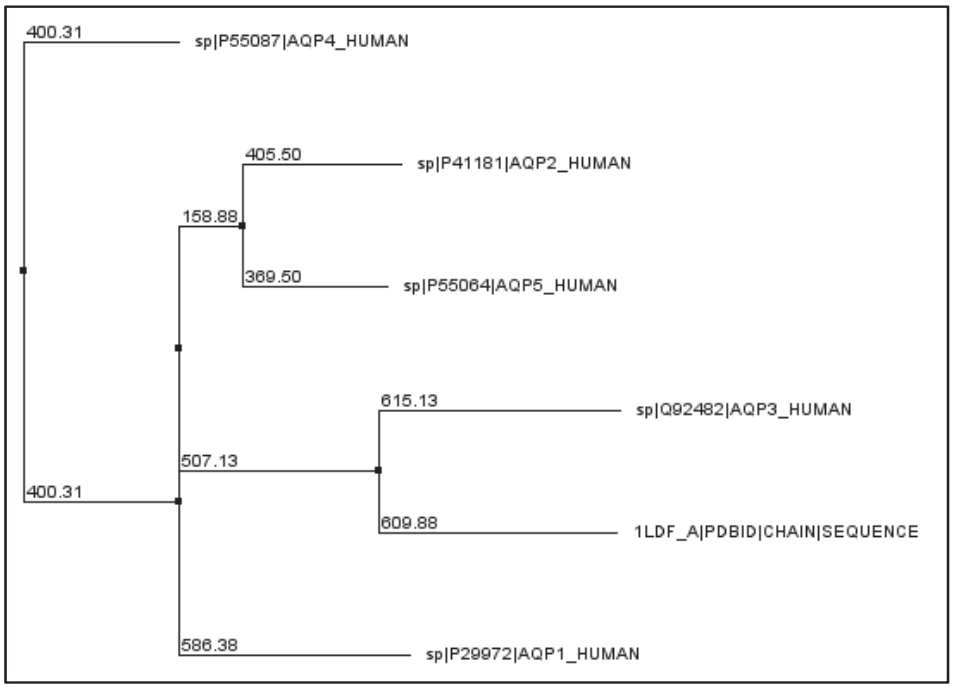


Fig. 4. The neighbor joining tree of aquaporin 1-5 and the glycerol channel of *E. coli* (PDB_ID 1ldf) indicates the sequence similarities of aquaporin-2,-3 and their modeling templates aquaporin-5 and 1ldf, respectively. The direct neighborhood of the involved proteins and their template structures point to the adequate reliability of the existing aquaporin-2 and aquaporin-3 models.

Mutations in aquaporin are correlated with NDI. For the detailed analyses of mutations in aquaporin-2, we concentrated on two well-defined point mutations. Characterization of D150E and G196D aquaporin-2 mutations are responsible for nephrogenic diabetes insipidus: importance of a mild phenotype. These two mutations were compared with the wild-type protein (aquaporin-2-wt) for functional activity (water flux analysis), protein maturation, and plasma membrane targeting. As shown by Guyon et al (2009) the D150E mutant induces an intermediate water flux compared to the aquaporin-2 –wt whereas the G196D mutant leads to no water flux. This observation is consistent with results from immunocytochemical experiments and Western blotting which indicate partial targeting of D105E in plasma membrane and complete sequestration of G196D within intracellular compartments. When coinjecting aquaporin-2-wt with mutants, no (aquaporin-2-wt + D150E) or partial (aquaporin-2-wt + G196D) reduction of water flux were observed compared with aquaporin-2-wt alone, whereas complete loss of function was found when both mutants were coinjected (Guyon, et al., 2009).

2.2 Model of V2 receptor

The V2 receptor belongs to the class A of G-protein-coupled receptors containing seven membrane spanning helices which are connected by extracellular and intracellular loops of varying length. The function of V2R is coupled to the G-protein activating adenyl cyclase (Barberis et al. 1998). If an agonist arginine vasopressin binds to V2R, the receptor becomes activated which leads to allosteric structural rearrangements. These structural changes then enable interactions with the cytosolic G-protein. The binding site of arginine vasopressin on

the V2 receptor is formed within the transmembrane helices II –VII (Slusarz et al. 2006). Regions between residues 88-96, 119-127, 284-291 and 311-317 contain most of the residues involved in binding. The selectivity of vasopressin was proposed to be connected with non-conserved residue Q180 whose carboxamide forms hydrogen bonds with carboxamide of Asn5 in the peptide. The stability of the hormone in the bound state is ensured by two hydrogen bonds between peptide backbone atoms of Tyr 2 and Asn 5. In general, hydrogen bonding and salt bridges were identified as the most important interactions contribution to the arginine vasopressin binding (Slusarz et al. 2006). Significant hydrophobic interactions were not detected.

A three-dimensional structural model of human V2 receptor (Fig. 5) was produced using I-TASSER protein structure modeling pipeline (Roy et al. 2010). I-TASSER builds protein models using multiple threading alignments on template structures and iterative assembly. The top three structural templates used in the structure prediction were PDB_ID 2ks9 (Substance-P receptor) with sequence identity of 19 %, PDB_ID 2rh1 (B2-adrenergic G protein-coupled receptor) with sequence identity of 22 % and PDB_ID 1l9h (bovine rhodopsin) with sequence identity of 18 % to VR2 receptor. The modeled structure was subjected first to conjugate gradient minimization and then MD simulation using the program NAMD2 (Phillips et al. 2005) and the CHARMM27 force-field (MacKerrel et al. 1998). The TIP3P solvent model represented the water molecules (Joergensen et al. 1983). Simulations assumed constant particle number, constant pressure and constant temperature (*NpT*) ensembles. *Langevin* dynamics was used to maintain constant temperature and pressure was controlled using a hybrid *Nose-Hoover Langevin* piston method. Extensive molecular dynamics simulations were done on the modeled structure in order to study its quality and structural stability. The average root-mean-square-deviation of the backbone atoms of the modeled structure during the simulation was found to be 2.7 Å. The model structure has seven helical segments: helix I 34 -64, helix II 73 – 101, helix III 109 – 142, helix IV 153-175, helix V 202-230, helix VI 248-296, helix VII 304-328.

Molecular dynamics (MD) simulation is one of the most common methods to study computationally protein function, conformational flexibility, and interactions. It is a technique to calculate the equilibrium and transport properties of a classical many-body system (Frenkel, 2002). In MD simulations, particles obey the laws of classical mechanics and the technique show how the system of particles evolves in time. The first MD simulation of a protein was done in the 1970s in vacuum for duration of 10ps (McCammon et al. 1977). Nowadays, the computational power allows simulation of about one million atoms, up to 100 Å in size and time scale up to 1 microsecond. Even single membrane proteins in the native lipid environment can be studied. Simulations of large biomolecular systems are becoming more feasible as demonstrated by the work on the MD-based structure prediction of the ribosome complex from *E. coli* (Villa, 2009), the simulation of the assembly of lipids and proteins into lipoprotein particles (Shih, 2007), the MD studies of viral capsid self-assembly (Freddolino, 2006; Miao, 2010) and vesicle fusion simulations (Kasson, 2010). MD simulations are used to gain information about the conformational changes of protein structure, *i.e.* sample the configuration space. In addition, MD simulations provide thermal averages of molecular properties such as the free energy change upon binding or atomic mean square fluctuation amplitudes. According to the ergodic hypothesis all microstates of a system are equally probable for a particle over a long period of time. Hence, the average of a process parameter over time and the average over the statistical ensemble are the equal. Simulation can be used to study the dynamics of a

system in detail by observing the conformational states that are accessible in a given temperature. *Ab initio* structure prediction starting from amino acid sequence of a protein using MD simulations is computationally feasible only for very small proteins, but simulations can be used to improve computational predicted protein structures obtained by homology modeling or fold recognition. The limitations of MD simulations are still relatively short time scale, inaccuracies in the description of physical interactions and the size limitation of the simulation system.

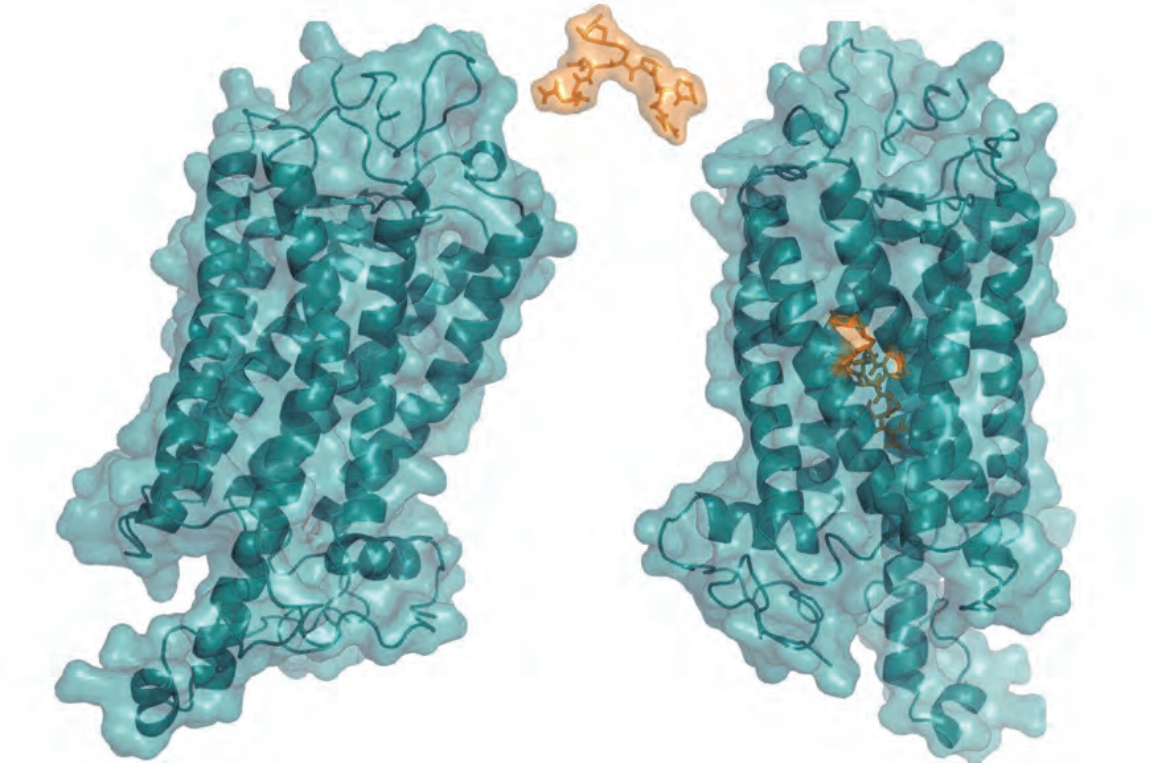


Fig. 5. Structure model of the V2 receptor (left). On the right side a result from Molecular Docking Server. The AVP hormone is highlighted in orange and is bound to the V2 receptor (right).

Mutation	Effect
A84D	This mutation not only affects receptor folding in such a way as to lead to its retention inside the intracellular compartments but, as expected, also has profound effects on its binding and coupling properties (pubmed_Id 10820167).
I130F	Functional analysis of I46K and I130F revealed reduced maximum agonist-induced cAMP responses as a result of an improper cell surface targeting (pubmed_Id 10770218,16006591)
P322S	P322S mutation of AVPR2 gene leads to a mild form of CNDI. (pubmed_Id 10026830,9402087)

Table 3. Overview of mutations in the V2 receptor and a short description of the biological effect. (More mutants in the appendix.)

In this work we address only a repertory. We do not analyses mutations cause constitutive activation of the receptor in this work (such as: Feldman et al, 2005). The table 3 shows the position and the molecular description of the investigated mutations of the V2 receptor be focused on this work.

3. SMFS – Stability and experiments

Atomic force microscopy (AFM) is mostly known for its imaging capabilities (Müller & Engel, 1999; Müller, et al, 1999; Seelert et.al, 2003). Recently, single-molecule force spectroscopy (SMFS) has proven to be a tool for detecting and locating inter- and intra-molecular forces on a single molecule level. SMFS experiments allow measuring the stability of membrane proteins and also probing the energy landscapes (Janshoff et al., 2000; Janovjak et al, 2004). In Fig. 6A a schematic representation of the force spectroscopy instrumentation is shown. Molecules with complex three-dimensional structures, such as proteins, can be unfolded in a controlled way. When transmembrane proteins are unfolded in force spectroscopy experiments, during continuous stretching of the molecule the applied force is

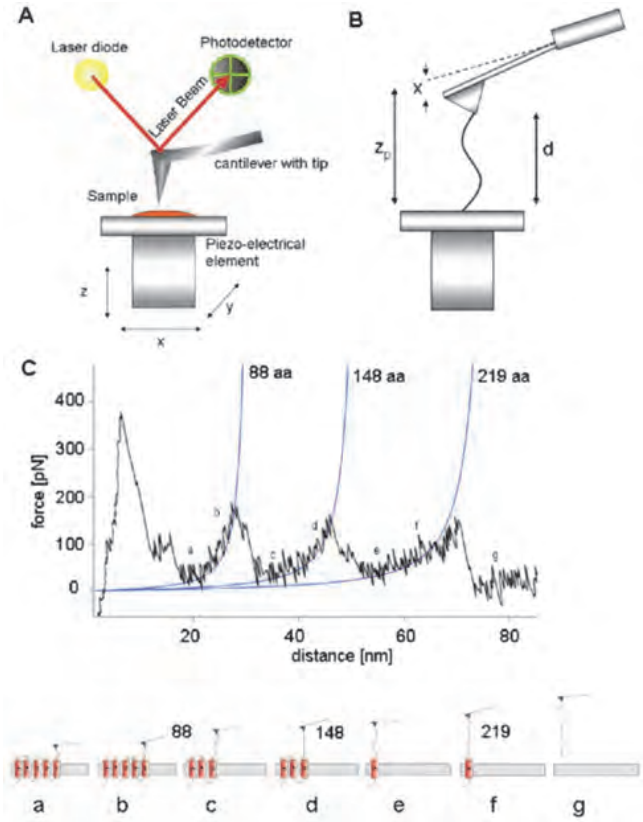


Fig. 6. A: Schematic representation of AFM. The sample is mounted on a piezo-electric element and scanned under a sharp tip attached to the cantilever. The voltage difference of the photodetector is proportional to the deflection of the cantilever.
B: Unfolding of a transmembrane protein. A single molecule is attached between the tip and the sample while the distance between tip and sample is continuously increased.
C: Typical spectrum obtained from an unfolding experiment of bR with the main peaks fitted by a hyperbolic function (WLC model) and correlated to the unfolding of secondary structure elements (cartoon at the bottom).

measured by the deflection of the cantilever and plotted against extension (tip-sample separation), yielding a characteristic force-distance curve (F-D curve) (see Fig. 6). From the analysis of single molecule force spectra it is possible to associate the peaks to individual stable structural segments within membrane proteins. For a given protein under study, the F-D curves exhibit certain patterns, which contain information about the strength and location of molecular forces established within the molecule, stable intermediates and reaction pathways, and the probability with which they occur. For membrane proteins the sequence of the unfolding peaks follow the amino acid sequence of the protein. Fitting each peak to a hyperbolic function, the worm-like chain model (WLC), gives the number of already unfolded amino acids (Rief et al. 1997).
Consequently, with the peaks and the predicted secondary structure, it is possible to associate the peaks to structural domains (see Fig. 7 and Fig. 8). Force curves show specific and unspecific interactions which lead to different unfolding pathways.
To draw biologically relevant conclusions on molecular interactions about how strong they are and where they occur, or whether they are independent or occur only in presence with other events, one must analyze many F-D curves by identical objective procedures. Thus, there is an increasing demand for data analysis techniques that offer fully automated processing of many datasets with identical analysis procedures. To discriminate force spectra showing specific and non-specific interactions and different unfolding/unbinding pathways, classification and pattern recognition algorithms are urgently needed (Marcico et al. 2007; Sapra et al., 2008).
One aim of the analyses of experimental data from SMFS measurements is the detection of possible unfolding pathways. Furthermore, we can identify different groups in hierarchical trees, which relate to different unfolding events. These events correspond to secondary structure elements and stabilized regions in the investigated protein.

3.1 SMFS experiments on aquaporin-1

Here, we work on data from SMFS experiments for human aquaporin-1 from the literature In the work of (Möller et al, 2003) 26 F-D curves were measured and manually aligned. The individual hAQP1 molecules were unfolded by pulling at their C-termini. The author created an overlay of all investigated curves (Fig. 7).

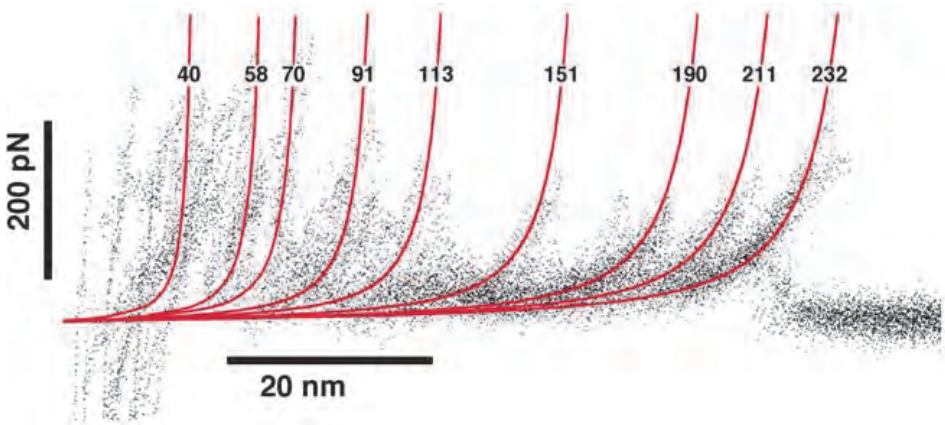


Fig. 7. Overlay of 26 F-D curves of the human aquaporin-2 and fitted using the WLC model (continuous curves). The numbers on the WLC fits indicate the contour lengths used to obtain the fit, in amino acids (Möller et al, 2003).

The next step is the mapping of the unfolding results of the known structure of aquaporin-1. This leads to a correlation of unfolding events and the secondary structure of the membrane protein. A possible description of all events is listed in table 4.

Contour length from WLC fits (aa)	Peak occurrence number/percent (n_{total} = 26)	Average force (pN)	Proposed potential barrier	Grey marker (Fig. 8)
40 ± 8	26 (100%)	206 ± 64	end of helix H6	1
58 ± 6	24 (92%)	157 ± 49	end of helix HE	2
70 ± 8	17 (65%)	125 ± 63	end of helix H5	3
91 ± 7	22 (85%)	156 ± 44	Helix H5	4
113 ± 7	13 (50%)	98 ± 54	end of helix H4	5
151 ± 5	20 (77%)	82 ± 53	Helix H3	6
190 ± 10	17 (65%)	98 ± 33	end of helix HB	7
211 ± 6	16 (62%)	77 ± 42	HelixH2	8
232 ± 5	26 (100%)	152 ± 62	HelixH1	9

Table 4. Contour lengths, peak occurrence, average forces, and positions of potential barriers in Aqp1 topology by SMFS experiments (additional link to Fig. 8 – topology).

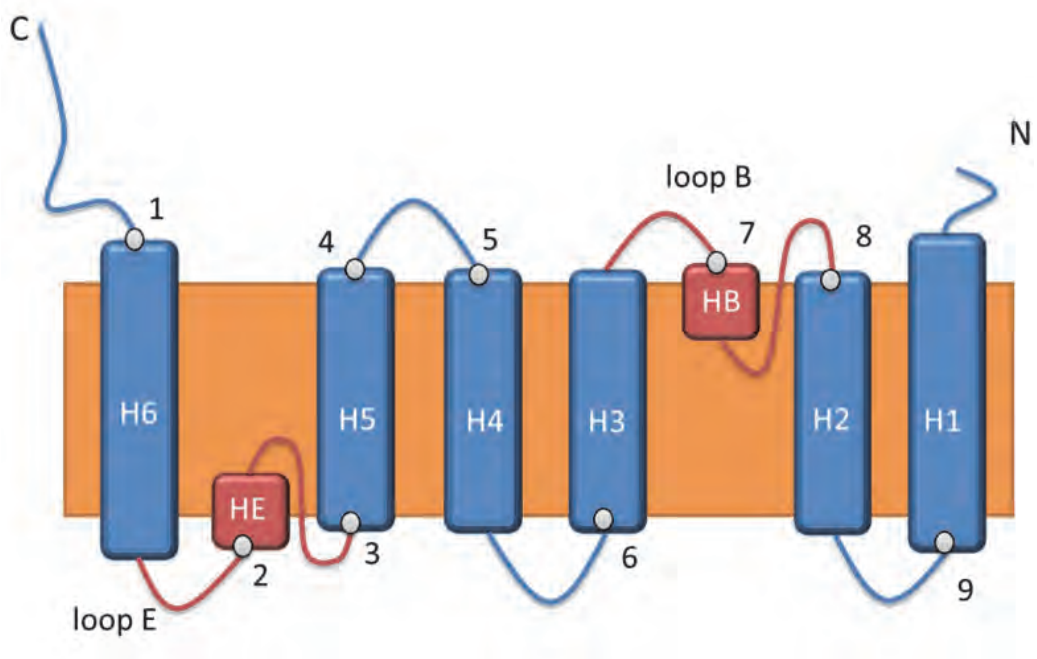


Fig. 8. Topology model of Aqp1: Shown are the secondary structure elements in the lipid bilayer, as described by the 3D structure. Numbers in ovals represent the numbers of proposed potential barrier of table 4.

Interesting are the two long loops, with formed helices in the transmembrane region. A view of the structure of aquaporin-1 quickly shows the role of both helices (HB and HE). The residues Asn 72 (part of HB) and Asn 192 (part of HE) arrange the immediate place of the pore (Fig. 8 and 9).

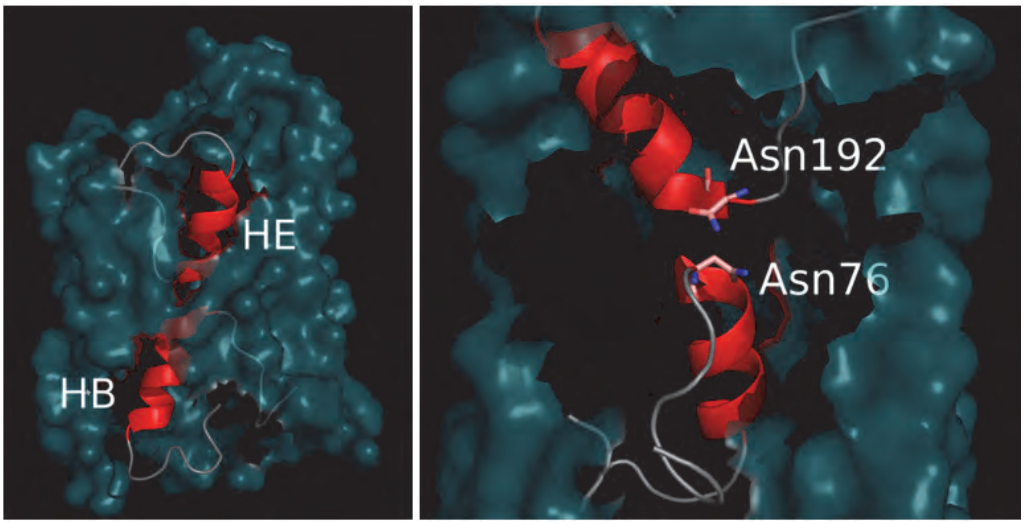


Fig. 9. Structure of aquaporin-1: Left side loop regions E and B (transmembrane helices HE and HB highlighted in red), right: Interface with conserved residues Asn 192 and Asn 76.

This important functional and structural feature corresponds to two unfolding events in aquaporin-1. The major force peak appeared within the noisy region but at a tip–membrane separation of about 20 nm. The WLC fit (Fig. 7 red line – 58 aa) showed an average contour length of 58±6 amino acids, while the rupture event exhibited an average force of 147±49 pN. According to the topology shown in Fig. 8, the extracellular end of helix HE is separated from the C-terminal end. Thus, this adhesion peak is likely to reflect the unfolding of HB. We can observe an analog situation for the loop B and the corresponding helix HB.
The force peaks found at a contour length of 190±10 amino acids (Fig.7) exhibit an average rupture force of 98±33 pN. This distance from the C-terminus corresponds to helix HB, which dips into the membrane from the cytoplasmic side and is only 11 amino acids long. Thus, this adhesion peak is likely to reflect the unfolding of HB.

3.2 Unfolding characteristic of aquaporin-2 and aquaporin-3

On the basis of known structures of aquaporin-1,-4 and -5 and the models of the aquaporin-2 and -3 we created a multiple structure alignment using the PDBeFold service of the EMBL-EBI. We clustered the resulting q-scores of all pair wise structural alignments by applying the UPGMA method (Unweighted Pair Group Method with Arithmetic Mean) to get a rooted tree (see Fig. 10, left). The inner node of the tree indicates that the known structures of aquaporin-5 and the glycerol channel (PDB_ID 1ldf) are almost identical in protein fold. The direct neighborhood of the aquaporin-1 structure and the models of aquaporin-2 and -3 give a strong hint for the unfolding characteristics of aquaporin-2. Due to their high structural similarity we postulate that aquaporin-2,-3 have a similar unfolding characteristic in comparison to aquaporin-1. The pair wise structural alignment of aquaporin-2 and -3 is shown in Fig. 10, right. The structure of aquaporin-4 shares a high similarity to the other structures.

4. Energy profiles – Stability and theory

A lot of tools and methods in the field of bioinformatics and structural biology are based on structure and/or sequence comparison. In this section we demonstrate a new method based on so called energy profiles for analyzing protein structure stability. Those profiles are

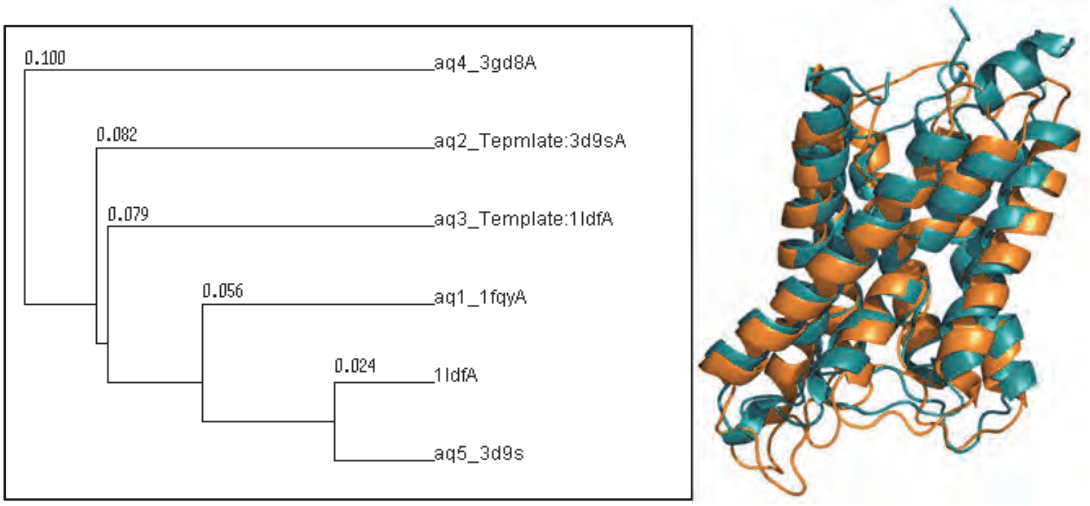


Fig. 10. Left: tree of the q-scores resulting from the PDBeFold service using Unweighted Pair Group Method with Arithmetic Mean (UPGMA) hierarchical clustering. Right: A structural alignment of aquaporin-1 structure (orange) and the model of aquaporin-2 (cyan).

calculated by coarse grained models. Based on the residue contacts in known protein structures, we calculated the potential for pair wise residue-residue-interactions. An energy profile is a schematic plot of the interaction energy of each residue as a function of the residue position in the sequence.

4.1 Theory of energy profiles

In this section, we show the theoretical aspects and calculation of so called protein energy profiles. The aspects explained in this section are essential in understanding the energy profile based methods we applied to the aquaporin proteins.

Energy profiles are derived by coarse-grained amino acid interaction models based on information of known protein structures. In general, the energy of any protein is given by equation (1), where e^*_{ij} acts as the interaction energy between two amino acids a_i and a_j. The function $f(r_{ij})$ quantifies e^*_{ij} by the Euclidean distance r_{ij}. The solvent interaction energy of an amino acid a_i is given by e'_{i0} and is relativized by expression $g(i)$, which describes the solvent accessibility state of a_i.

$$E = \sum_{<ij>} e_{ij} f(r_{ij}) + \sum_{i} e'_{i0} g(i) \tag{1}$$

Based on (1), we designed a coarse grained interaction scheme, which uses the C_α and C_β coordinates of the amino acids. Furthermore we redefined $g(i)$ and $f(r_{ij})$. Instead of using a continuous space, $f(r_{ij})$ acts as Boolean function. That means, depending on r_{ij}, amino acid a_i is either interacting with amino acid a_j or it is not. Based on the work of (Dressel et al. 2007; Wertz & Scheraga, 1978) we defined a cut-off threshold for r_{ij} of 8Å. That leads to the equation (2).

$$f(r_{ij}) = \begin{cases} 1 & \text{if } r_{ij} \text{ is} \le 8\text{Å}, \\ 0 & \text{else.} \end{cases} \tag{2}$$

ÅFurthermore we introduced an amino acid specific inside/outside-property which reflects the orientation of the amino acid side chains with respect to the center of mass of the neighboring residues and was defined in the following way:
A residue is declared as inside, if

$$\left|\vec{C}_\alpha - \vec{c}\right| < 5 \vee \left(\vec{C}_\alpha - \vec{C}_\beta\right)\left(\vec{C}_\alpha - \vec{c}\right) < 0 \tag{3}$$

$\vec{C}_{\alpha/\beta}$ are the vectors of the $C_{\alpha/\beta}$ atoms, $\vec{c}$ is the center of mass of all amino acids in a surrounding sphere with a radius of 5Å. For determining the center of mass only C_α atoms are taken into account. Using this property the inverse Boltzmann equation can be applied to calculate the energy of each amino acid a_i in the protein structure by (4).

$$e_i = -k_B T \ln\left(\frac{n_{(i,in)}}{n_{(i,out)}}\right) \tag{4}$$

The parameters $n_{(i,in)}$ and $n_{(i,out)}$ are equal to the number of inside and outside occurrences of amino acid a_i, respectively. These parameters are derived by known globular and membrane protein structures. In our coarse grained model, the interaction energy e_{ij} between two amino acids a_i and a_j is equal to the summation of e_i and e_j. Finally, let S be a set of amino acids, let $k = |S|$ and a_i is defined as the observed amino acid. For each $a_j \in S$ is $r_{ij} \leq 8$Å. Then the total energy E_i of a_i equals (5). By iterating over all amino acids in a protein structure the total energy for each amino acid can be calculated and the energy profile is generated.

$$E_i = \sum_{j=1}^{k}\left(e_i + e_j\right) = \sum_{i=1}^{k}\left(-k_B T \ln\left(\frac{n_{(i,in)}}{n_{(i,out)}}\right) - k_B T \ln\left(\frac{n_{(j,in)}}{n_{(j,out)}}\right)\right) \tag{5}$$

Additionally, it needs to be said that we discard further solvent interaction calculation (seen in the second summation in equation 1) because these information is modeled by the amino acid specific inside/outside-property. In addition, we declared T as constant which leads to discarding the constant $-k_B T$ in the energy profile calculation. Thus the energies, which result by our model, are arbitrary unit entities [a.u.] and are direct proportional to energies given in [J] or [kcal·mol^{-1}].
In conclusion, by calculating the total energy of an amino acid in a protein structure, physicochemical and structural information are abstracted to one single value.
The relation of amino acid stability and amino acid energy is explainable by the folding of the protein and its energy landscape. As one of the last steps in protein biosynthesis the polypeptide folds into the native protein structure state spontaneously which is equal to the proteins most stable fold. This process can be described as a function of the loss of the Helmholtz free energy within an amino acid interaction energy state. Commonly a folded protein in its stable state holds the minimized amount of free energy. The energy profile is a transformation of the energy landscape of the protein at the point of minimized free energy, which leads to the conclusion that the energy value of an amino acid a_i given by an energy profile is a transformation of the stability of the amino acid a_i in the structure. Figure 11 illustrates the resulting energy profile (right) of the structure model of aquaporin-2 (left).

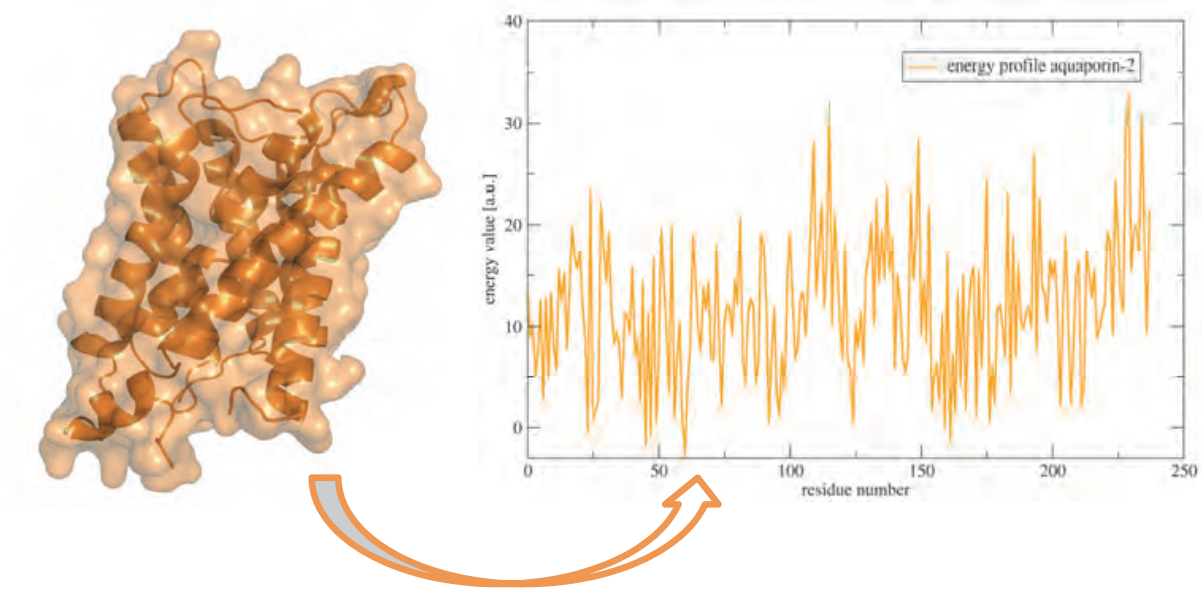


Fig. 11. Left: Structure model of aquaporin-2 and the corresponding energy profile (right) of this model.

4.2 Energy profiles analyses of the investigated aquaporins

On the basis of the so-called energy profiles we can compare the structures and models of all investigated aquaporins, well-defined mutations and the influences of mutations for the stability of the aquaporins.

4.2.1 Energy profiles of the investigated aquaporins

For calculating the energy profiles we used the structures given in table 1 and table 2. To evaluate the energy profiles we checked the already existing aquaporin models concerning their reliability. As shown, the best matching structures were used as template structures for homology modeling. On the level of energy profiles we can confirm our hypothesis that all investigated aquaporins have the same stability characteristics. For this purpose we created a multiple energy profile alignment (MEPAL). We adapted standard algorithms in clustering and deriving consensus profiles and energy conservation. Figure 12 shows the MEPAL for all of the involved aquaporins in **Diabetes Insipidus** and the human aquaporin-1.

The MEPAL method is based on classical multiple sequence alignment algorithms using modified scoring functions optimized for energy profile comparison. The tree (see Fig. 13) is calculated by applying the UPGMA clustering method to the pair wise distance scores which are calculated by the MEPAL algorithm. Furthermore, the graphical alignment output (Figure 12) consists of three parts. The upper row shows the energetically aligned energy profiles represented by the amino acids of the protein sequence which are colored depending on their energy. The greater the energy of an amino acid the greater is the red color content. The middle row shows the consensus profile. In the consensus profile, each energy at position *i* is derived by calculating the pair wise distance sores of all aligned energies at position *i*. The energy with the lowest average distance is representing the consensus profile at position *i*. Finally, the bottom row shows the conservations at each alignment position. Each conservation value is calculated by the sum of pair wise energy distances and is normalized by the number of aligned profiles (Gusfield 1993, Gusfield 1997).

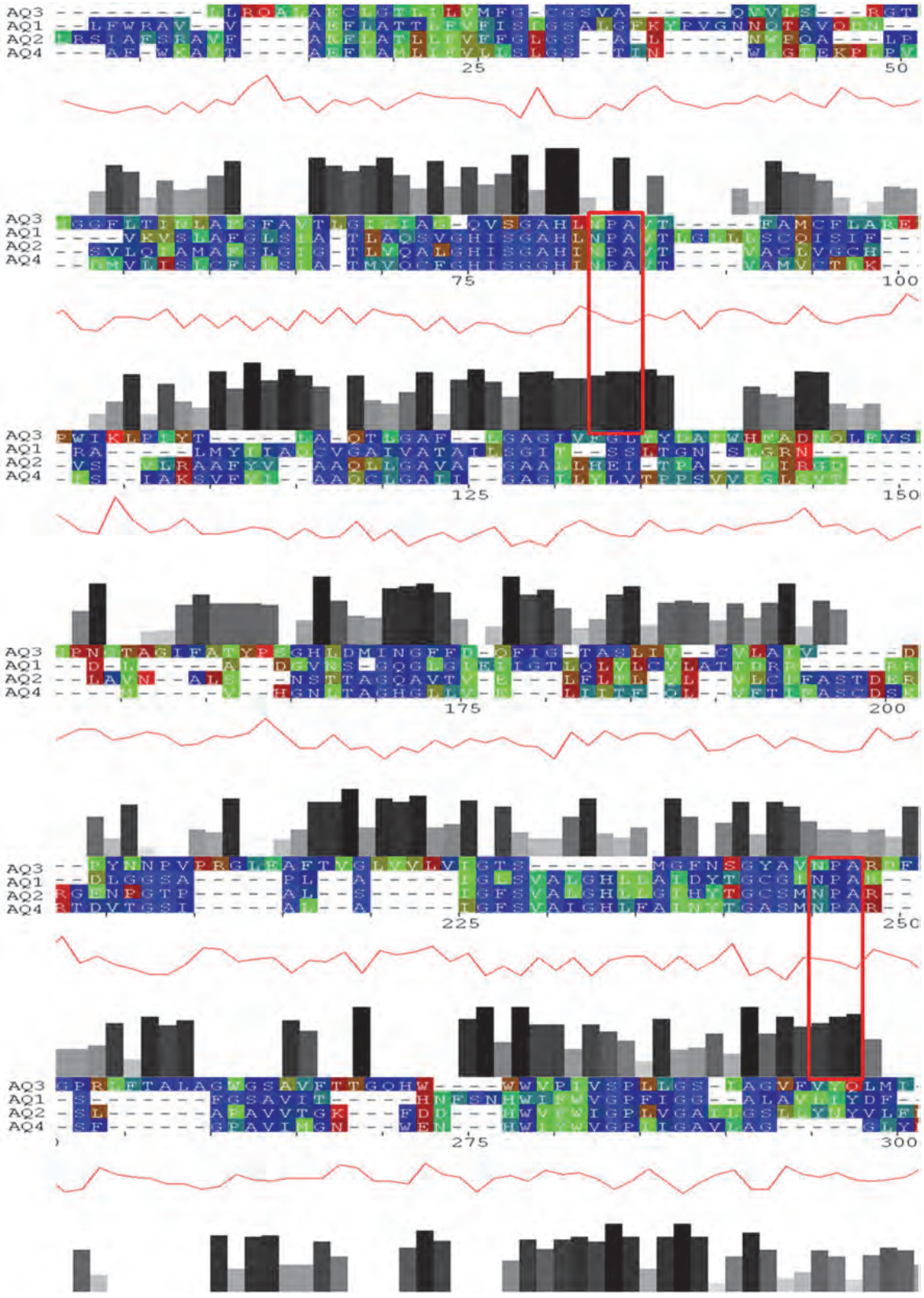


Fig. 12. MEPAL output for the energy profile alignment of aquaporin-1, -2,- 3 and -4. The Asn-Pro-Ala motifs are highlighted by red boxes.

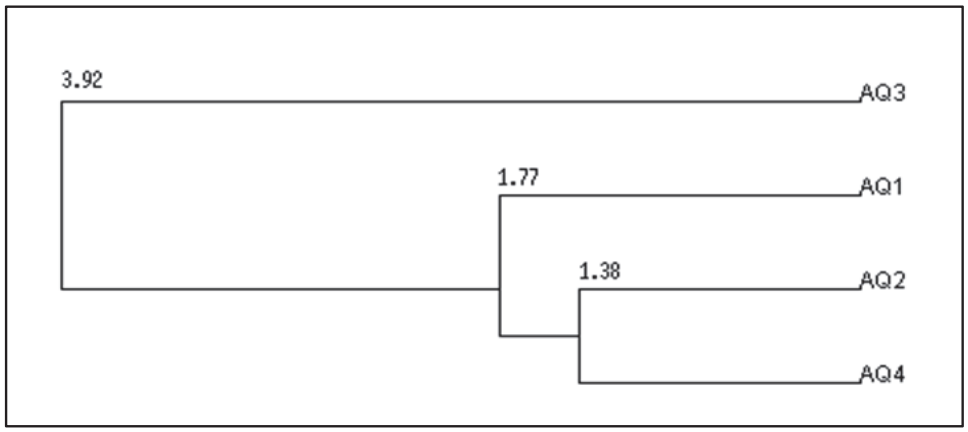
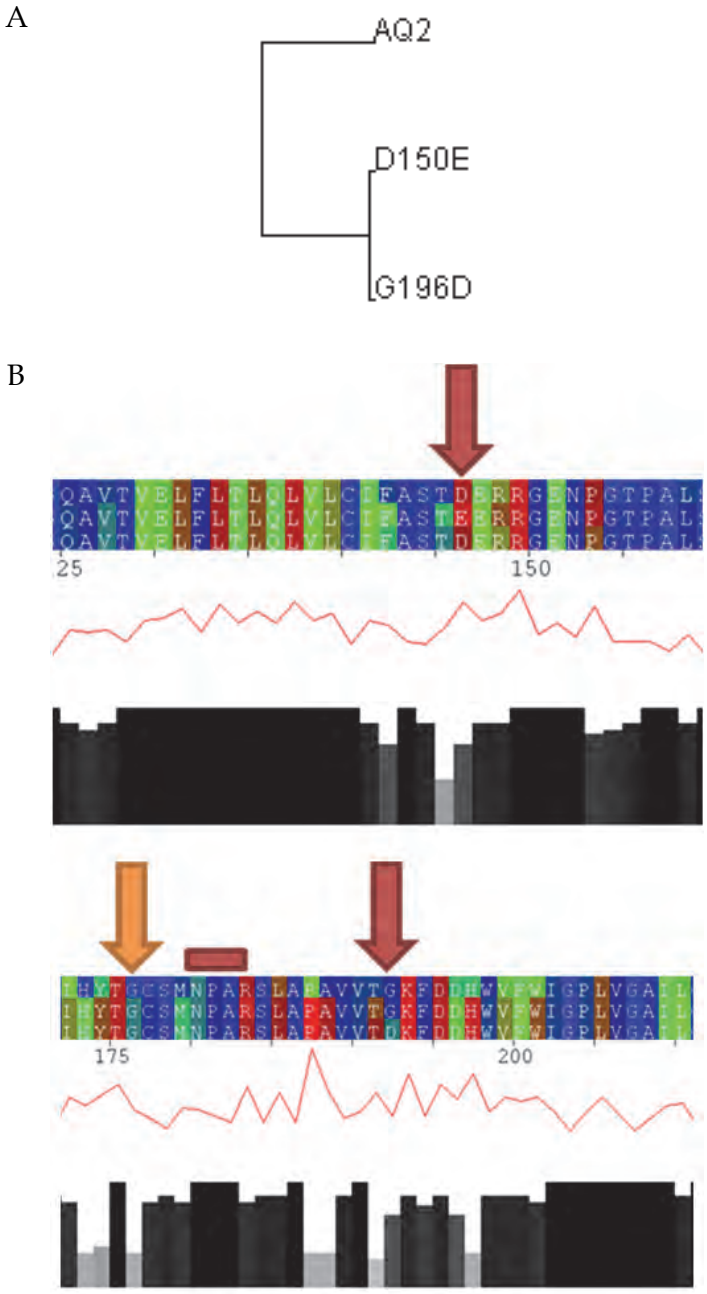


Fig. 13. UPGMA tree, based on energy profiles of aquaporin-1, -2, -3 and -4.

The energy profile alignment based UPGMA Tree, which was calculated by MEPAL indicates high similarities between the energy profiles of aquaporin-1,-2 and -4 and is seen in figure 13. The distance of 3.92 between aquaporin-3 and the other structures corresponds to significant similarity. The graphical output of the MEPAL illustrates several highly energetically conserved amino acids and regions. Two of these conserved regions correspond to the opposite orientated Asn-Pro-Ala motifs in helix HB and helix HE (see section 2.1). These two motifs are highlighted by red boxes in Fig. 12. The energetic conservation of these motifs and their surrounding amino acids confirms the importance of these residues in water transport in aquaporin. Additionally the residues Gly 188, Phe 56, Cys 189, Ile 191 and His 180, which are involved in water transport as well, show differences in sequential and energetic conservation. In detail the conserved amino acids Gly 188 and Phe 56 show slight divergences or no divergences at all concerning their calculated energy. Cys 189 and Ile 191 show no conservation in aquaporin 1-4; but these changes have no effect on the level of energy profiles. Missing in aquaporin-3, His 180 shows sequential and energetic conservation in the other aquaporins. We postulate that these slight divergences do not affect the water flux significantly.
A further point of interest lies in Arg 195. This residue is conserved in all four proteins but varies energetically. These divergences arise from conformational changes of the residue and the structural environment. Based on the facts we referred to in section 2.1, we postulate that these divergences between aquaporin 1-4 lead to a change in the secondary free-energy barrier influencing the transport selectivity and the water flux. It also needs to be said that the significant differences in the energy profile progression between aquaporin-3 and the other structures (see Fig. 13) might result by the less reliable aquaporin-3 model. Despite these divergences, we can confirm our postulated similarities concerning the unfolding characteristics of the aquaporins involved in **Diabetes Insipidus**.

4.2.2 Energy profiles and stability of the mutants of aquaporin-2

For comparison on the level of energy profiles we generated aquaporin-2 models with the two mutations: D150E and G196D. Based on all three models we calculated the energy profiles and created a MEPAL. The results lead to a distance tree and can be discussed on the level of aligned energy profiles (Fig.14).



A: The UPGMA tree of the resulting distance scores of the energy profiles calculated by MEPAL. The inner node indicates that these point mutations lead similar energy profiles.
B: The MEPAL output of the investigated energy profiles. The point mutations induce various energetic changes which are highlighted by arrows. The red rectangle illustrates the Asn-Pro-Ala motif.

Fig. 14. Results from the analyses of the energy profile of aquaporin-2 and the investigated mutants.

The energy profile based UPGMA tree (see Fig. 14, A) indicates strong similarities between the energy profiles of the two modelled aquaporin-2 mutants. This leads to the conclusion that both mutants induce the same energetic, structural and functional changes. While both mutations led to energetic variations in the entire energy profiles we focused our discussion on the mutations sites (see Fig. 14, B - red arrows). It needs to be said that because of the

modelling procedure and the energy profile calculation the resulting energy profile covers not all amino acids of the mutated sequence. In this case, this leads to an index indention of 3 amino acids. The mutation D150E (see Fig. 14, B - at the top) induces an energetically increase of the two surrounding residues decreasing the energetic conservation at these positions. Interestingly, in this region the mutation G196D induces almost the same energetic increase as D150E. At the mutation site of the modelled G196D variant (see Fig. 14, B - at the bottom), the mutation induces slight energetic divergences in the region where the mutation site is located. Furthermore, in this region the G196D mutation leads to the same energetic variations as the D150E mutation. This observation can be confirmed at nearly all positions in the energy profiles of the modelled aquaporin mutations. Interestingly, both mutations do not affect the energetic conservation of the Asn-Pro-Ala motif (highlighted by a red rectangle in Fig. 14, B at the bottom).
Additionally, we point to the energetic changes of Gly 188 (highlighted by an orange arrow in 14 B at the bottom). As mentioned in section 2.1 this residue is involved in water transport. Both mutations lead to an energetically increase of Gly 188 and reduce the energetic conservation in these three investigated energy profiles. Thus, we postulate that the mutations D150E and G196D affect the transecullar water transport.

4.3 Energy profiles analyses of the V2 receptor

For investigating energetic influences, binding capabilities and the effect of mutations in the V2 receptor we generated a V2 receptor model by molecular modeling. This model was used to calculate the energy profile of this receptor. Furthermore, we used the Molecular Docking Server to process a docking simulation of the V2 receptor model and the arginine vasopressin hormone. The structure of the model and its hormone in docked state was used for further analyses. Both models are illustrated in Fig. 5 in subsection 2.2. We calculated the energy profile of the V2 receptor model in docked state to detect energetic divergences induced by conformational changes and the hormone itself. Thus, AVP is an oligopeptide it can be integrated into the energy profile calculation in the way it is explained in section 4.1.
To detect docking induced energetic changes both derived energy profiles were aligned using the MEPAL method. The alignment revealed energetic divergences in the surrounding of the amino acids Ala 84 (Fig. 15, A), Ile 130 (Fig. 15, B) and Pro 322 (Fig. 15, C). This leads to the conclusion that these amino acids are involved in hormone binding. Mutations of these amino acids are well described in literature and are causing a loss in functionality and hormone affinity (see table 3). Our novel energy profile based approach brought more evidence to these described mutations.

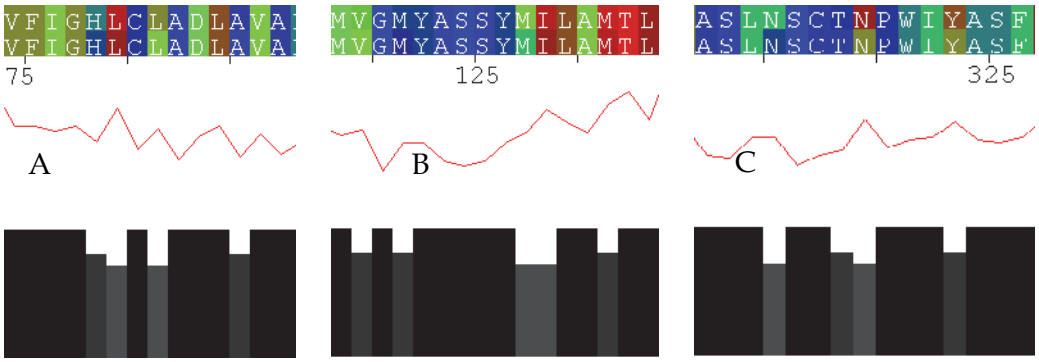


Fig. 15. MEPAL output for the energy profile alignment of fragments of V2 receptor and the complex

5. Discussion of the stability of the investigated membrane proteins

The antidiuretic hormone, ADH, also called vasopressin and arginine-vasopressin, is a nanopeptide (nine amino acids) synthesized in the hypothalamus, transported to and stored in the posterior lobe of the pituitary gland which releases it into the blood circulation. It has antidiuretic and vasopressor actions. The effects of vasopressin result from stimulation of V1 and V2 receptors, V1 mainly responsible for vasoconstriction, V2 for the antidiuretic effect V1 receptors are coupled by G-protein to phospholipase C. V2 receptors are coupled by G-protein to adenylcyclase. Its activation elicits an increase in cAMP which, via protein kinases, induces the activation of aqueous channel aquaporin-2 -mainly located in the renal collecting duct. Under the influence of vasopressin aquaporin-2 migrates from the cytoplasm to the apical membrane. In **nephrogenic diabetes insipidus** there are aquaporin-2 alterations.
Aquaporin-3 is constitutively expressed in the basolateral membrane of the cell. When water floods into the cell through aquaporin-2 channels, it can rapidly exit the cell through the aquaporin-3, 4 channels and flow into blood.
We investigated the stability of membrane proteins on the basis of experimental and theoretical assumptions. Membrane proteins play essential roles in cellular processes. These mutations cause structural instabilities in a transmembrane protein leading it to unfold or missfold in an alternative conformation. By the structural comparison of the aquaporin-1, investigated by SMFS, with the involved proteins and protein models, we can postulate that aquaporin-2,-3 and -4 exhibit similarities in unfolding characteristics. This assumption can be confirmed by our theoretical approach. These theoretical methods are based on the so called energy profile calculation which is explained in this work. On the basis of stability analyses and the application of energy profile based methods we were able to enforce evidence for water flux reduction induced by well described mutations. Furthermore the correlation of residue conservation and energetic conservation of amino acids involved in water transport was detected. Especially the role of the two conserved helices HB and HE could be detected and described on the basis of experimental and theoretical methods.
For analyzing the V2 receptor we derived a structure model by molecular modeling and processed AVP docking simulations. As a main part we focused on selected point mutations and their influences in hormone affinity. Thus, we were able to enforce evidence described in literature.

6. Acknowledgment

This project was funded by the Free State of Saxony and the University of Applied Sciences Mittweida. The authors thank Daniel Stockmann for helpful discussions, motivations and powerful programming.

7. Appendix

In the preparation of the book chapter we collected the literature for well-defined mutations in the V2 receptor. We don't claim to have a complete list or all description. On the basis of this list we compared our results in the context of the docking model of the complex V2 receptor and the AVP hormone.

Mutation	Effect
N22Q	N-linked glycosylation at asparagine 22, Mutagenesis of asparagine 22 to glutamine abolished N-linked glycosylation of the V2 receptor (N22Q-V2R), without altering its function or level of expression. (pubmed_Id 10362843)
L44F	the mutant L44F and the in vitro mutant S167A were expressed in their mature form at wild-type levels (pubmed_Id 8863826)
L44P	mutants L44P, W164S, S167L, and S167T lacked complex glycosylation and were expressed at low levels, mutants misfolded (pubmed_Id 8863826,16006591)
S45C	Strong beta-catenin (CTNNB1) expression in the tumor cells and identified a heterozygote missense Ser45Cys mutation of exon 3 of CTNNB1 (pubmed_Id 19294427).
I46K	Functional analysis of I46K and I130F revealed reduced maximum agonist-induced cAMP responses as a result of an improper cell surface targeting (pubmed_Id 10770218)
L62P	core-glycosylated mutants L62P and V226E were excluded from lysosomes (pubmed_Id 18048502)
A84D	this mutation not only affects receptor folding in such a way as to lead to its retention inside the intracellular compartments but, as expected, also has profound effects on its binding and coupling properties (pubmed_Id 10820167).
A98P	the cell-surface expressions of mutant receptors were totally (A98P and L274P) (pubmed_Id 17371330)
W99R	Mutation of a tryptophan located at the beginning of the first extracellular loop (W99R) that greatly impaired the binding properties of the receptor and had a minor effect on its intracellular routing (pubmed_Id 10820167).
F105V	the F105V mutation is delivered to the cell surface and displayed an unchanged maximum cAMP response, but impaired ligand binding abilities of F105V were reflected in a shifted concentration-response curve toward higher vasopressin concentrations. As the extracellularly located F105 is highly conserved among the vasopressin/oxytocin receptor family, functional analysis of this residue implicates an important role in high affinity agonist binding. (pubmed_Id 10770218)
R113W	The cell-surface expressions of mutant receptors were totally (A98P and L274P) or partially (R113W) absent. V2R-R113W, -G201D, and -T204N were expressed in the ER and in the basolateral membrane as immature, high-mannose glycosylated, and mature complex-glycosylated proteins. The immature forms of V2R-R113W and -T204N, but not V2R-G201D, were rapidly degraded. The mature forms varied extensively in their stability and were degraded by only lysosomes (V2R-T204N and wild-type V2R) or lysosomes and proteasomes (V2R-G201D, -R113W). (pubmed_Id 17371330,16006591,7984150,10770218)

I130F	Functional analysis of I46K and I130F revealed reduced maximum agonist-induced cAMP responses as a result of an improper cell surface targeting (pubmed_Id 10770218,16006591)
R137C	(R137C) in the second intracellular loop, which has been associated with constitutive activation of the AVPR2. In conclusion, adults with intermittent, severe hyponatraemia may have a constitutively activating mutation in the AVPR2 with resultant nephrogenic syndrome of inappropriate antidiuresis. R137C gain-of-function mutation was detected by means of mutation analysis of the V2R gene. (pubmed_Id 18753429,18622631,16843086,19179480,17229917)
R137L	V2R-R137L mutant interacts with beta-arrestins in an agonist-independent manner resulting in dynamin-dependent internalization. V2R-R137L mutant traffic considerably more efficiently to the plasma membrane than V2R-R137H, identifying this as a potentially important mutation-dependent difference affecting V2R function. (pubmed_Id 19179480,16843086)
R143P	R143P and delta V278 mutants are retained within the cytoplasmic compartment. (pubmed_Id 7560098)
S167A	The mutant S167A was functionally active, (pubmed_Id 8863826)
R181C G185C	loss of receptor function (pubmed_Id 15841479)
G201D	the complex-glycosylated mutant G201D were partially located in lysosomes, G201D was expressed in the ER and in the basolateral membrane as immature, high-mannose glycosylated, and mature complex-glycosylated proteins. (pubmed_Id 18048502,16006591)
R202C	R202C mutant reaches the cell surface, a simple binding impairment at the cell surface (pubmed_Id 7560098)
T204N	degraded by only lysosomes, T204N was expressed in the ER and in the basolateral membrane as immature, high-mannose glycosylated, and mature complex-glycosylated protein (pubmed_Id 16006591)
Y205C Y205F Y205H	-for Y205C the lack of a Tyr residue at position 205 is responsible for the abolished receptor function rather than the formation of a disastrous second disulfide bond. Y205C mutant was almost inactive. -Analysis of the intermolecular interaction of the Tyr-205 hydrogen group by molecular modeling showed that Tyr-205 was located in transmembrane domain (TM) 5, and that its hydroxy group formed a hydrogen bond with Leu-169 main-chain =O located in TM 4. The mutation of Tyr-205 to phenylalanine would cause loss of this hydrogen bond and decrease or change the interaction between these TM coils, thus affecting the ability of AVP to bind to the receptor. According to this molecular model of AVPR2, the Y205F mutation would cause nephrogenic diabetes insipidus. - the loss of receptor function of Y205H, NDI-causing mutation Y205H which affects a codon frequently found to be mutated to Cys in NDI patients. (pubmed_Id 15841479,11026555, 17216256,)
V206D	stimulation of the V206D mutant increased the cAMP accumulation only slightly, V206D was mainly expressed in the endoplasmic reticulum (ER) as immature proteins. (pubmed_Id 11026555,16006591)

F287L	F287L mutant in COS-7 cells revealed significant dysfunction and accumulate intracellular cyclic adenosine monophosphate in response to AVP hormone stimulation. (pubmed_Id 11916004)
P322S	P322S mutation of AVPR2 gene leads to a mild form of CNDI. (pubmed_Id 10026830,9402087)
S363A	The S363A mutation that confers recycling to the V2R did not alter its interaction with arrestins. (pubmed_Id 11353798)

Table appendix: Overview of mutations in the V2 receptor and a short description of functional and structural influences.

8. References

Robertson GL (1995) Diabetes insipidus. *Endocrinol Metab Clin North Am* 24: 549-572.

Ananthakrishnan S (2009) Diabetes insipidus in pregnancy: etiology, evaluation, and management *Endocr Pract* 15: 377-382.

Krysiak, R.; Kobielusz-Gembala, I. et al (2010) Recurrent pregnancy-induced diabetes insipidus in a woman with hemochromatosis Endocrine Journal, online-ISSN: 1348-4540

Deen PM, Verdijk MA, Knoers NV, Wieringa B, Monnens LA, van Os CH, van Oost BA.Requirement of human renal water channel aquaporin-2 for vasopressin-dependent concentration of urine. *Science* 1994; 264: 92–95.

Mulders SM, Bichet DG, Rijss JPL, Kamsteeg EJ, Arthus MF, Lonergan M, Fujiwara M,

Morgan K, Leijendekker R, van der Sluijs P, van Os CH, Deen PMT. An aquaporin-2 water channel mutant which causes autosomal dominant nephrogenic diabetes insipidus is retained in the Golgi complex. *J Clin Invest* 1998; 102: 57–66.

Van den Ouweland AMW, Dreesen JCFM, Verdijk M, Knoers NVAM, Monnens LAH, Rocchi M, VanOost BA. Mutations in the vasopressin type-2 receptor gene (Avpr2) associated with nephrogenic diabetes-insipidus. *Nat Genet* 1992; 2: 99–102.

Rosenthal W, Seibold A, Antaramian A, Lonergan M, Arthus MF, Hendy GN, Birnbaumer M, Bichet DG. Molecular-identification of the gene responsible for congenital nephrogenic diabetes-insipidus. *Nature* 1992; 359: 233–235

Los, E. L .; Deen, P. M. T; Robben, J. H. Potential of Nonpeptide (Ant)agonists to Rescue Vasopressin V2 Receptor Mutants for the Treatment of X-linked Nephrogenic Diabetes Insipidus. *Journal of Neuroendocrinology* 2010; 22: 393–399

Fujiwara TM, Bichet DG. Molecular biology of hereditary diabetes insipidus. J AmSoc Nephrol. 2005;16:2836–46.

Robben JH, Knoers NVAM, Deen PMT. Cell biological aspects of the vasopressin type-2 receptor and aquaporin 2 water channel in nephrogenic diabetes insipidus. Am J Physiol Renal Physiol 2006; 291: F257–F270.

Hardy, C.; Khanim, F.; Torres, R. et al. Clinical and Molecular Genetic Analysis of 19 Wolfram Syndrome Kindreds Demonstrating a Wide Spectrum of Mutations in WFS1. Am. *J. Hum. Genet*. 1999; 65:1279–1290.

Wolfram, D.J. and Wagener, H.P. (1938) Diabetes mellitus and simple optic atrophy among siblings: report on four cases. *Mayo Clinic Proc.*,13, 715–718.

Swift, M. and Swift, R.G. (2001) Psychiatric disorders and mutations at the Wolfram syndrome locus. *Biol. Psychiatry*, 47, 787–793.

Strom, T.; Hörtnagel, K.; Hofmann, S. Diabetes insipidus, diabetes mellitus, optic atrophy and deafness (DIDMOAD) caused by mutations in a novel gene (wolframin) coding for a predicted transmembrane protein. *Human Molecular Genetics*, 1998, Vol. 7, No. 13: 2021–2028

Inoue, H., Tanizawa, Y., Wasson, J., Behn, P., Kalidas, K., Bernal-Mizrachi, E., Mueckler, M.,
Marshall, H., Donis-Keller, H., Crock, P. et al. (1998) A gene encoding a transmembrane protein is mutated in patients with diabetes mellitus and optic atrophy (Wolfram syndrome). *Nat. Genet.*, 20, 143–148.

Strom, T.M., Hoetnagael, K., Hofmann, S., Gekeler, F., Scharfe, C., Rabl, W., Gerbitz, K.D. and Meitinger, T. (1998) Diabetes insipidus, diabetes mellitus, optic atrophy and deafness (DIDMOAD) caused by mutations in a novel gene (wilframin) coding for a predicted transmembrane protein. *Hum. Mol. Genet.*, 7, 2021–2028.

Cryns, K., Sivakumaran, T.A., Van den Ouweland, J.M., Pennings, R.J., Cremers, C.W., Flothmann, K., Young, T.L., Smith, R.J., Lesperance, M.M. and Van Camp, G. (2003)
Mutational spectrum of the WFS1 gene in Wolfram syndrome, nonsyndromic hearing impairment, diabetes mellitus, and psychiatric disease. *Hum. Mut.*, 22, 275–287.

Takeda, K., Inoue, H., Tanizawa, Y., Matsuzaki, Y., Oba, J., Watanabe, Y., Shinoda, K. and Oka, Y. (2001) WFS1 (Wolfram syndrome 1) gene product: predominant subcellular localization to endoplasmic reticulum in cultured cells and neuronal expression in rat brain. *Hum. Mol. Genet.*, 10, 477–484.

Hofmann, S., Philbrook, C., Gerbitz, K.D. and Bauer, M.F. (2003) Wolfram syndrome: structural and functional analyses of mutant and wild-type wolframin, the WFS1 gene product. *Hum. Mol. Genet.*, 12, 2003–2012.

King LS, Kozono D, Agre P. From structure to disease: the evolving tale ofaquaporin biology. *Nat Rev Mol Cell Biol*. 2004 5(9):687-98. Review.

Pollard TD, Earnshaw, WC. (2007), *Cell Biology*, Springer, ISBN 978-3-8274-1861-6, Heidelberg

Guyon C, Lussier Y, Bissonnette P, Leduc-Nadeau A, Lonergan M, Arthus MF, Perez RB, Tiulpakov A, Lapointe JY, Bichet DG. Characterization of D150E and G196D aquaporin-2 mutations responsible for nephrogenic diabetes insipidus: importance of a mild phenotype. *Am J Physiol Renal Physiol*. 2009;297(2):F489-98.

Ambrish Roy, Alper Kucukural and Yang Zhang, TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc*. 2010 April; 5(4): 725–738

D. Frenkel and B.J. Smit. Understanding Molecular Simulation. From Algorithms to Applications. Elsevier, 2002.

Jr. A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, and B. Prodhom. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 1998.

J. A. McCammon and M. Karplus. Internal motions of antibody molecules. *Nature*, 268(5622):765{766, Aug 1977.

Peter L. Freddolino, Anton S. Arkhipov, Steven B. Larson, Alexander McPherson, and Klaus Schulten. Molecular dynamics simulations of the complete satellite tobacco mosaic virus.*Structure*, 14:437-449, 2006.

Assembly of lipids and proteins into lipoprotein particles. Amy Y. Shih, Anton Arkhipov, Peter L. Freddolino, Stephen G. Sligar, and Klaus Schulten. Assembly of lipids and proteins into lipoprotein particles *Journal of Physical Chemistry B*, 111:11095-11104, 2007.

Yinglong Miao, Peter J. Ortoleva, Viral structural transition mechanisms revealed by multiscale molecular dynamics/order parameter extrapolation simulation. *Biopolymers* Volume 93, Issue 1, pages 61–73, January 2010Peter M. Kasson, Erik Lindahl, and Vijay S. Pande. Atomic-resolution simulations predict a transition state for *vesicle* fusion defined by contact of a few lipid. *PLoS Computational Biology*, *2010* June; 6(6): e1000829.

Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys* 1983, *79*, 926-935.

James C. Phillips, Rosemary Braun Wei Wang James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, Klaus Schulten, Scalable molecular dynamics with NAMD

Barberis C, Mouillac B, Durroux T.,*J Endocrinol.* 1998 6(2):223-9. Structural bases of vasopressin/oxytocin receptor function.

Slusarz MJ, Giełdoń A, Slusarz R, Ciarkowski J. Analysis of interactions responsible for vasopressin binding to human neurohypophyseal hormone receptors-molecular dynamics study of the activated receptor-vasopressin-G(alpha) systems. *J Pept Sci*. 2006 Mar;12(3):180-9.

Feldman BJ, Rosenthal SM, Vargas GA, Fenwick RG, Huang EA, Matsuda-Abedini M, Lustig RH, Mathias RS, Portale AA, Miller WL, Gitelman SE.: Nephrogenic syndrome of inappropriate antidiuresis., N Engl J Med. 2005 352(18):1884-90.

Müller, D.J., Engel, A.: Voltage and pH-induced channel closure of porin OmpF visualized by atomic force microscopy. J. Mol. Biol. 285, 1347–1351 (1999)

Müller, D.J., Sass, H.J., Muller, S.A., Buldt, G., Engel, A.: Surface structures of native bacteriorhodopsin depend on the molecular packing arrangement in the membrane. *J. Mol.Biol*. 285, 1903–1909 (1999)

Seelert, H., Dencher, N.A., Muller, D.J.: Fourteen protomers compose the oligomer III of the proton-rotor in spinach chloroplast ATP synthase. *J. Mol. Biol*. 333, 337–344 (2003)

Janshoff, A., Neitzert, M., Oberdorfer, Y., Fuchs, H.: Force spectroscopy of molecular systems-single molecule spectroscopy of polymers and biomolecules. *Angew Chem. Int. Ed.* Engl. 39(18), 3212–3237 (2000)

Janovjak, H., Struckmeier, J., Hubain, M., Kedrov, A., Kessler, M., Muller, D.J.: Probing the energy landscape of the membrane protein bR. *Structure* 12(5), 871–879 (2004)

Rief, M., Gautel, M., Oesterhelt, F., Fernandez, J.M., Gaub, H.E.: Reversible unfolding of individual titin immunoglobulin domains by afm. *Science* 276(5315), 1109–1112 (1997)

Marsico A, Labudde D, Sapra T, Muller DJ, Schroeder M. A novel pattern recognition algorithm to classify membrane protein unfolding pathways with high-throughput single-molecule force spectroscopy. *Bioinformatics*. 2007

Sapra KT, Balasubramanian GP, Labudde D, Bowie JU, Muller DJ. Point mutations in membrane proteins reshape energy landscape and populate different unfolding pathways. *J Mol Biol*. 2008 Feb 29;376(4):1076-90.

Möller C, Fotiadis D, Suda K, Engel A, Kessler M, Müller DJ. Determining molecular forces that stabilize human aquaporin-1. *J Struct Biol*. 2003 Jun;142(3):369-78.

Chen H, Wu Y, Voth GA. Origins of proton transport behavior from selectivity domain mutations of the aquaporin-1 channel., *Biophys J.* 2006;90(10):L73-5.

de Groot, B. L., T. Frigato, V. Helms, and H. Grubmüller. 2003. The mechanism of proton exclusion in the aquaporin-1 water channel. *J. Mol. Biol.* 333:279–293.

Chakrabarti, N., E. Tajkhorshid, B. Roux, and R. Pomes. 2004. Molecular basis of proton blockage in aquaporins. *Structure.* 12:65–74.

Chakrabarti, N., B. Roux, and R. Pome`s. 2004. Structural determinants of proton blockage in aquaporins. *J. Mol. Biol.* 343:493–510.

Ilan, B., E. Tajkhorshid, K. Schulten, and G. A. Voth. 2004. The mechanism of proton exclusion in aquaporin channel. *Proteins*. 55: 223–228.

Dressel, F., Tuukkanen, A., Schroeder M., and Labudde, D., Understanding of SMFS barriers by means of energy profiles, *Proc. GCB*, 2007

Wertz, D. H. and Scheraga, H. A., Influence of water on protein structure. An Analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule. *Macromolecules*. 1978 Jan-Feb;11(1):9-15.

Gusfield D., Efficient Methods for Multiple Sequence Alignment with Guaranteed Error Bounds, *Bulletin of Mathematical Biology*, Vol. 55, p. 141-154, 1993

Gusfield D., Algorithms on Strings, Trees and Sequences, Cambridge University Press, 1997

2.1.2 Membrane Protein Stability Analyses by Means of Protein Energy Profiles in Case of Nephrogenic Diabetes Insipidus

This paper summarizes the findings discussed in the book chapter presented in the last section. However, the paper emphasizes the theoretical aspects of the analyses. The formal presentation has been rewritten and re-evaluation of the employed membrane protein structures is discussed.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2012, Article ID 790281, 11 pages
doi:10.1155/2012/790281

Research Article

Membrane Protein Stability Analyses by Means of Protein Energy Profiles in Case of Nephrogenic Diabetes Insipidus

Florian Heinke and Dirk Labudde

Department of Mathematics, Natural and Computer Sciences, Hochschule Mittweida, University of Applied Sciences, Technikumplatz 17, 09648 Mittweida, Germany

Correspondence should be addressed to Dirk Labudde, dirk.labudde@hs-mittweida.de

Received 24 November 2011; Accepted 4 January 2012

Academic Editor: Silvina Matysiak

Diabetes insipidus (DI) is a rare endocrine, inheritable disorder with low incidences in an estimated one per 25,000–30,000 live births. This disease is characterized by polyuria and compensatory polydypsia. The diverse underlying causes of DI can be central defects, in which no functional arginine vasopressin (AVP) is released from the pituitary or can be a result of defects in the kidney (nephrogenic DI, NDI). NDI is a disorder in which patients are unable to concentrate their urine despite the presence of AVP. This antidiuretic hormone regulates the process of water reabsorption from the prourine that is formed in the kidney. It binds to its type-2 receptor (V2R) in the kidney induces a cAMP-driven cascade, which leads to the insertion of aquaporin-2 water channels into the apical membrane. Mutations in the genes of V2R and aquaporin-2 often lead to NDI. We investigated a structure model of V2R in its bound and unbound state regarding protein stability using a novel protein energy profile approach. Furthermore, these techniques were applied to the wild-type and selected mutations of aquaporin-2. We show that our results correspond well to experimental water ux analysis, which confirms the applicability of our theoretical approach to equivalent problems.

1. Introduction

Membrane proteins play important roles in many biological processes. Although the total number of known membrane protein structures has increased from 337 to 1515 structures within the last eight years, the high degree of redundancy and the average quality of these structures reduce the overall condition of structural data significantly [1]. At the moment, only 398 nonredundant membrane protein structures are available by protein structure databases, such as the Protein Data Bank (PDB) or the Protein Data Bank of Transmembrane Proteins [2, 3]. Hence, little is known about membrane proteins. To investigate membrane protein structure and misfolding, other approaches, such as small-molecular-force spectroscopy, have been applied and developed [4]. Mutation-induced membrane protein structure misfoldings are causes of many human diseases, that is, *diabetes insipidus*, *hereditary deafness*, *retinitis pigmentosa*, *cystic fibrosis*, *familial hypercholesterolaemia,* and so on [4–7].

Diabetes insipidus (DI) is characterized by polyuria (a daily output of 15–20 L of highly dilute urine) and compensatory polydypsia. In the general population, it is assessed on one case per 25,000–30,000 people [8–10]. Symptoms of DI in newborn infants are irritability, poor feeding, poor weight gain, and dehydration. DI can be differentiated in two classes. First, *central diabetes insipidus* (CDI) is caused by central defects, in which no or an insufficient amount of functional arginine vasopressin (AVP) is released from the pituitary. In contrary, defects in the kidney could cause *nephrogenic diabetes insipidus* (NDI). Four different types of NDI concerning causes and inheritance are known [11–14]:

(i) acquired NDI, it can originate as a side effect of long, surpassing drug taking (i.e., lithium);

(ii) autosomal recessive inheritable NDI, caused by mutations in AQP2 gene which encodes aquaporin-2;

(iii) dominant inheritable NDI, caused by mutations in AQP2 gene which encodes aquaporin-2;

(iv) X-linked inheritable NDI, caused by mutations in AVPR2 gene which encodes the AVP type-2 receptor (V2R).

The X-linked inheritable variant of NDI is a disorder in which a person affected is unable to concentrate urine in the kidney despite the presence of AVP. This nanopeptide (10 amino acids) acts as an antidiuretic hormone. It binds to V2R as an agonist and induces a cAMP-driven cascade which, as one result, leads to the insertion of aquaporin-2 in the apical collecting duct membrane. As an α-helical membrane water channel, the fusion of aquaporin-2 with the cell membrane increases the permeability of apical plasma membranes to water. Thus, water can pass through the apical membrane and leads to the prourine concentration equilibrium. Misfolded V2R mutants trapped in the endoplasmatic reticulum are the main cause for the origin of the X-linked NDI variant. Usually V2R fuses with the basolateral membrane where it is able to bind to AVP. Furthermore, inserted mutants are usually not able to bind with AVP. Thus, trapped V2R mutants or normally inserted but not-functional V2R mutations anticipate the induction of aquaporin-2 insertion which results to polyuria and diuresis. Autosomal recessive or dominant inherited mutations in the AQP2 gene lead to the misfolding of aquaporin-2 and, hence, the insertion of functional aquaporin-2 water channels. This results in the typical NDI symptoms elucidated above as well [15, 16]. Mutations in V2R and aquaporin-2 cause structural instabilities. The analysis of these instabilities plays an important part concerning the understanding process of membrane protein mutation-induced diseases, especially in *diabetes insipidus*.

In this paper, we demonstrate a novel approach for membrane protein stability analysis based on protein energy profiles. The concept of protein energy profiles is a novel coarse-grained model for transforming structural and chemical protein properties to one-dimensional energy representations. A protein energy profile can be calculated by any given protein structure within less than half a second making it valuable for the fast analysis of structure-function relationships. This approach is explained closer in the Material and Methods section. Its application to a structure model of V2R in bound and unbound state to AVP will be elucidated in detail. Additionally, the energy profile based membrane protein analysis is applied to the structure model of the wild-type of aquaporin-2 and selected mutant structure models. Finally, significant differences in the energy characteristics and correlations to experimental data will be discussed in detail.

2. Materials and Methods

2.1. Description of the Investigated Proteins

2.1.1. V2 Vasopressin Receptor (V2R). The V2 vasopressin receptor belongs to the class A of G-protein coupled receptors. It contains seven membrane spanning helices which are connected by extracellular and intracellular loops, respectively. The binding of V2R to agonist AVP induces the activation of the protein leading to allosteric structural rearrangements [17]. Once V2R is activated, it is able to interact with the cytosolic G-protein activating adenyl cyclase which triggers a cAMP-driven cascade. As a result aquaporin-2 is inserted in the apical membrane [15, 16]. The binding site of AVP in V2R is located within transmembrane helices II-IV, where the residues 88–96, 119–127, 284–291, and 311–317 are mainly involved in binding [18, 19].

Since there is no known protein structure of V2R, a three-dimensional structure model of V2R was produced using the I-TASSER protein structure modeling pipeline [20]. Basically I-TASSER builds protein models using iterative assembling procedures and multiple threading alignments based on template structures. In Table 1, the PDB_IDs, corresponding biological descriptions and sequence identities to V2R of the employed template structures, are given. Gradient minimization of the modeled structure was produced by means of NAMD2 [21]. Further MD simulation was applied to the model by using CHARMM27 force-field [22]. To study the overall model quality and structural stability, additional MD simulations were performed. The average RMSD of the C_α-backbone of the structure model was found to be 2.7 Å.

2.1.2. Aquaporin-2. Aquaporins provide highly permeable pores for water to cross membranes. Four identical subunits form a stable tetramer spanned through the plasma membrane. Each subunit consists of seven helices which form a pore with 3 Å in diameter. The selectivity for water is achieved mainly by the two asparagine residues 76 and 192 (human aquaporin-1 numbering) [23]. Further selective residues are His180, Gly188, Cys189, Gly190, Ile191, Arg195, Phe56. An illustration of the residues in the aquaporin-1 structure which are involved in water binding is given in Figure 1 [23, 24]. Furthermore, all aquaporins exhibit two highly conserved Asn-ProAla motifs which are located in two opposite meeting α-helices in all known aquaporin structures. This indicates the high conservation of this structural feature and its importance in water transport activity. It is shown that these two α-helices form a bipolar electric field changing the water molecule orientation and preventing protons to pass the channel. Further molecular simulation studies have revealed a secondary free energy barrier which is induced by Phe56, His180, and Arg195. It is located about 8 Å apart from the primary bipolar electric field at the extracellular side of the protein. As one result of these two bipolar electric fields, a constriction region is formed which allows only a single water molecule to pass the end of the pore. MD simulations of Arg195 mutants revealed the correspondence of the stability of this secondary bipolar electric field to Arg195 and, hence, the influences on water selectivity of the protein [25–27]. Although no high resolution structure of aquaporin-2 is given, all discovered features in known aquaporin structures and knowledge can be assigned reliably to aquaporin-2 due to the mostly high conservation of these residues in all known human aquaporins.

For the protein-energy-profile-based analysis, a structure model is necessary. Therefore, the most reliable structure

Table 1: Overview of the used template structures applied in modeling the V2R receptor structure.

PDB_ID	Description	Sequence identity to template [%]
2ks9	Substance-P receptor	19
2rh1	B2-adrenergic G-protein coupled receptor	22
1l9h	Bovine rhodopsin	18

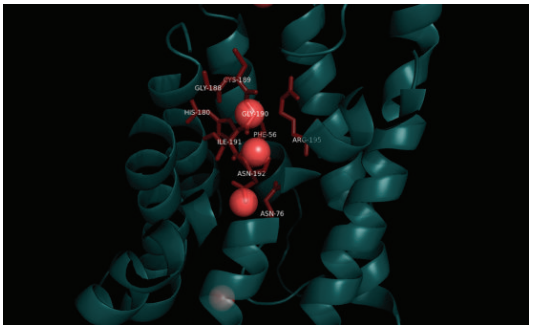


Figure 1: Illustration of the water binding network in aquaporin-1 (PDB_ID 1fqy). Due to high residue conservation, knowledge gained from known aquaporin structures can be assigned to aquaporin-2 reliably. Since there is no high resolution structure of aquaporin-2, this gained knowledge sheds light on the structure-function relationship in aquaporin-2.

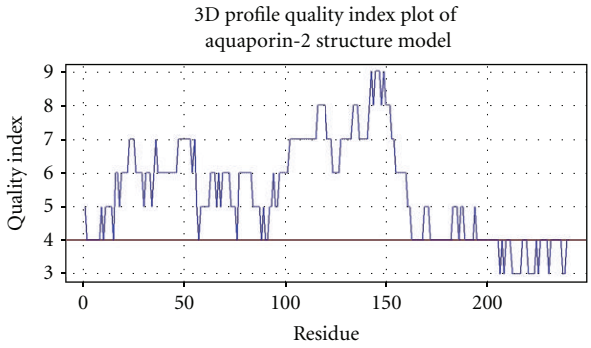


Figure 2: The quality index plot of the aquaporin-2 structure model produced by VADAR. The structure model of aquaporin-2, which is necessary for protein-energy-profile-based analysis, was reevaluated using the structure analysis tool VADAR [29]. One evaluation criterion given by VADAR is the quality index per residue. It indicates weak spots in the model, for example, side chain misfolding, stereochemical overlaps, insufficient atom packing, and so on. Here, the plot points to a high structure model quality.

model was retrieved from the ModBase database [28]. This model was produced using the high resolution structure of aquaporin-5 as the modeling template (PDB_ID 3d9s). Both proteins share a sequence identity of 68%. The used model exhibits a high coverage of 93% to the template structure. To reevaluate the quality of the model, the protein structure analysis tool VADAR (version 1.8, see [29]) was applied. One evaluation criterion is the quality index which summarizes side chain misfoldings, stereochemical overlaps, and insufficient atom packing for each residue. Quality indexes below 4 are reported as low quality. The quality index plot of aquaporin-2 is shown in Figure 2. As illustrated, the average quality index of all residues of the aquaporin-2 model points to a high quality level with only a few weak spots. Thus, the model can be applied to the protein-energy-profile-based approach.

2.2. Theory of Protein Energy Profiles. Since the fundamental work of Anfinsen, which states that the native protein conformation is determined by the sum of amino acid residue interactions and, thus, by the amino acid sequence [30], many coarse- and fine-grained, all atom models describing residue-residue interactions were developed and adapted. They are based either on first principles approaches using physics laws or make use of knowledge of existing experimentally derived structures and statistical analysis. The latter approaches, the so called knowledge-based energy potentials (KBPs), assume that free energy functions describe the behavior of a protein structure and that, according to Boltzmann's principle, the low-energy states are observed with high frequency. KBPs differ in their level of description of system details: ranging from all-atom models and potentials to simplified coarse-grained models. The physics-based approaches to predict protein structure use molecular mechanics force fields which describe proteins at atomic detail and energy terms containing contributions from electrostatic and van der Waals interactions as well as covalent bonding of the polypeptide chain [31–33, 32]. However, such atomic detail simulations are only feasible for rather small proteins usually shorter than 150 amino acids.

The coarse-grained model for calculating protein energy profiles, which is described here, belongs to the KBPs approaches. Its basis stems from [34–36]. It is similar to the approach described in the work of Eisenberg et al. [37]. Eisenbergs approach transforms the three-dimensional structure of the protein to a one-dimensional representation by analyzing the structural environment of each residue in the structure. This environment is described by the buried surface of the residue side chain and the surface of the side chain which is exposed to polar atoms as well as the local secondary structure of the residue. In contrast, the approach described here analyses the environment of the investigated residue too but approximates its energy by reverting to pseudoenergies derived by statistical physics. These pseudoenergies are based on the tendency of each amino acid for being either buried or exposed to the solvent. Applying the Boltzmann principles to these tendencies, the pseudoenergy of each amino acid can be approximated. For instance, an exposed cysteine holds a higher energy as expected since cysteine is usually buried inside the protein structure [38].

Based on [34, 35], we defined an inside/outside property for generating amino acid buriedness distributions and, hence, allowing the pseudoenergy approximation. Let i denote one of the 20 canonical amino acids. $n_{\text{in},i}$ and $n_{\text{out},i}$ describe the absolute frequency of the amino acid i being assigned as "inside" and "outside" by the inside/outside property, respectively. The inside/outside-property is defined as

$$f(i) = \begin{cases} n_{\text{in},i}{+}{+}, & ||C_{\alpha,i} - c|| < 5\,\text{Å} \vee \left(C_{\alpha,i} - C_{\beta,i}\right)(C_{\alpha,i} - c) < 0, \\ n_{\text{out},i}{+}{+}, & \text{else}, \end{cases} \tag{1}$$

where c denotes the center of mass of all C_α atoms within a 5 Å sphere surrounding i. In general, the statistics can be calculated by any given set of proteins but redundancy and physiochemical properties need to be taken into account. For instance, the statistics of α-helical membrane proteins differ significantly from statistics derived by globular proteins exclusively. Exchanging statistics or calculating on the basis of a rather inappropriate protein structure set would lead to false conclusions. Here, the statistics were derived by employing this property to 342 nonredundant α-helical membrane proteins. The list of these protein structures can be found at the Protein Data Bank of Transmembrane Proteins [1]. Applying the inverse Boltzmann principle, the pseudoenergy e_i of i can be approximated as follows:

$$e_i = -k_B T \ln\left(\frac{n_{\text{in},i}}{n_{\text{out},i}}\right). \tag{2}$$

Since k_B and T are declared as constants in this model, both can be omitted from the calculation:

$$e_i^* = -\ln\left(\frac{n_{\text{in},i}}{n_{\text{out},i}}\right). \tag{3}$$

The energy of the pairwise interactions of i to other residues corresponds to the environment of i and the environments composition inside the structure [39]. Thus, the expected tendency value P of the observed environment composition correlates with the interaction energy of i. P can be approximated by the derived amino acid distributions:

$$P_{k\in\text{Env}} = \prod_{k\in\text{Env}} p_k = \prod_{k\in\text{Env}} \left(\frac{n_{\text{in},k}}{n_{\text{out},k}}\right), \qquad \ln P_{k\in\text{Env}} = \sum_{k\in\text{Env}} \ln\left(\frac{n_{\text{in},k}}{n_{\text{out},k}}\right). \tag{4}$$

Thus, according to the Boltzmann principle, the energy of the environment E_{Env} is defined as

$$E_{\text{Env}} = -\ln P_{k\in\text{Env}}, \tag{5}$$

and, hence,

$$E_i = -|\text{Env}| \ln\left(\frac{n_{\text{in},i}}{n_{\text{out},i}}\right) - \sum_{k\in\text{Env}} \ln\left(\frac{n_{\text{in},k}}{n_{\text{out},k}}\right). \tag{6}$$

The environment was defined by a contact function $g(i, j)$ adapted from [36] which is denoted as

$$g(i,j) = \begin{cases} 1, & \left|\left|C_{\alpha,i} - C_{\alpha,j}\right|\right| \le 8\,\text{Å}, \\ 0, & \text{else}, \end{cases} \tag{7}$$

Finally, the total energy of i is

$$E_i^* = \sum_{j\in S\setminus i} \left[g(i,j)\left(e_i^* + e_j^*\right)\right], \tag{8}$$

where S defines a given protein structure. By omitting $k_B T$ in the model, the resulting E_i^* are given in arbitrary unit entities [a.u.] and are direct proportional to energies listed in $[J]$ or $[\text{kcal mol}^{-1}]$. The protein energy profile of S corresponds to the n-tupel of all E_i^*.

Similar to the approach discussed in the work of Eisenberg et al. [37], energy profiles can be aligned by means of dynamic programming. Therefore, an energy-energy scoring function was implemented. It is derived by distances between power-equal intervals of the gaussian integral of the energy distribution. For scoring two energies, each energy is assigned to its interval in the gaussian integral. The distance between both integrals corresponds to the pairwise energy score. This scoring is used for aligning two given energy profiles A and B by dynamic programming, like the Needleman-Wunsch algorithm [40] or the Smith-Waterman algorithm [41]. The estimation of alignment significance is permitted by weighting the resulting score x_{r} to the best possible score x_{opt} and the average permutation score $\overline{x}_{\text{p}}$ which is derived by permuting and realigning the given energy profiles. As discussed in [42], this weighted score is called distance score (dScore) and is defined as

$$\text{dScore}(x_{\text{r}}) = -\log\left(\frac{x_{\text{r}} - \overline{x}_{\text{p}}}{x_{\text{opt}} - \overline{x}_{\text{p}}}\right) \tag{9}$$

with

$$x_{\text{opt}}(A,B) = \frac{\delta(|A| + |B|)}{2}. \tag{10}$$

Here, δ denotes the best possible pairwise energy score. In general, significant energy profile alignments correspond to dScores of less than 2.5 bans. The alignment of two identical energy profiles corresponds to a dScore of 0 bans (Figure 3).

2.3. Correlations of Energy Profiles to Structure and Amino Acid Sequence. The relation of amino acid stability and amino acid energy is explainable by the folding of the protein and its energy landscape. The process of folding can be described as a function of the loss of the free Helmholtz energy within an amino acid interaction energy state. Commonly, a folded protein in its stable state holds the minimized amount of free energy [43]. The energy profile is a transformation of the energy landscape of the protein at the point of minimized free energy. This leads to the conclusion that the energy value of an amino acid i given by an energy profile is a transformation of the stability of the amino acid i in the structure.

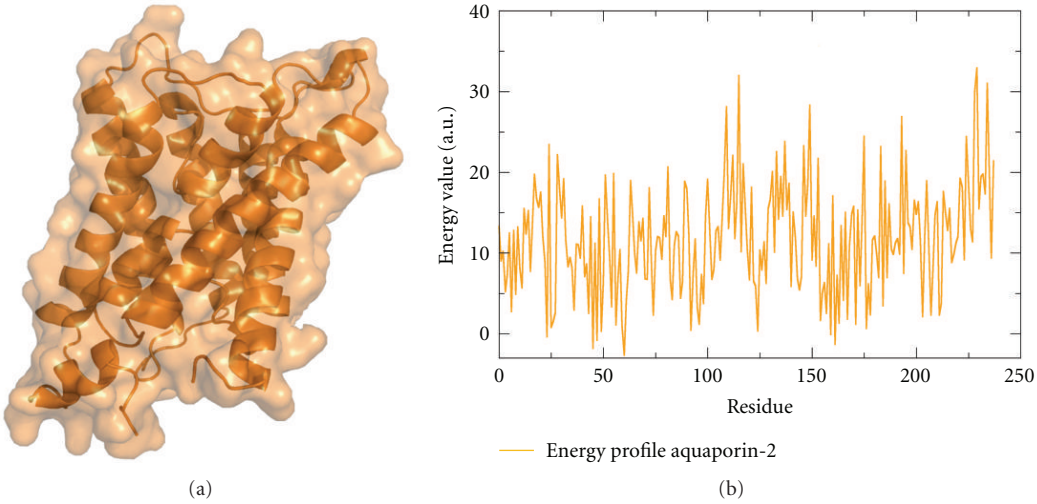


Figure 3: The structure model of aquaporin-2 (a) and its corresponding energy profile (b).

To investigate the correlation of energy profiles to structure and amino acid sequence, seven protein structures (PDB_IDs: 1a1w, 1a3h, 1amm, 1b1j, 1bhe, 1o8w, 3gbs) which share no similarity in structure, sequence, or function were subjected to the PDBeFold service [44] for searching identical and similar protein structures in the Protein Data Bank. Overall, PDBeFold detected 653 significant hits. For each hit, the sequence identity and structure alignment scores (QScores) were saved. Two identical protein structures afford a QScore of 1.0. In contrast, two dissimilar proteins share a QScore of 0.0. Afterwards, the protein energy profiles of the query protein structures and the corresponding hits were calculated and aligned. The Spearman correlation coefficients of the resulting dScores and saved QScores as well as the sequence identities were calculated. The resulting Spearman correlation coefficients are

(i) QScore to dScore: $\rho_{\text{QScore, dScore}} = -0.91$,

(ii) sequence identity to dScore: $\rho_{\text{seqId, dScore}} = -0.92$,

(iii) sequence identity to QScore: $\rho_{\text{seqId, QScore}} = 0.94$.

The corresponding scatter plots are shown in Figure 4. The high correlation of energy profile similarity to sequence identity and structure similarity leads to the conclusion that energy profile similarity achieves at least the same correlation as sequence identity to protein structure. This implies transitivity of the energy profile of a protein to its amino acid sequence and structure.

Additionally, the correspondence of secondary structure elements to their holding energy was investigated. Therefore, energy profiles of a nonredundant set of 342 α-helical membrane proteins were fragmented and labeled according to the secondary structure elements. The fragment length was five residues. 600 fragments were chosen randomly from the entire set of fragments and clustered using neural gas [46]. During the clustering process, the labels were ignored and served only for evaluating the clustering performance. The clustering was evaluated using normalized mutual information (NMI)[47]. This procedure was repeated in 100 iterations to verify the resulting values. Here, the average NMI was found to be at a low level of significance at 0.06 which means that there exists almost no correlation of secondary structures to their holding energy in α-helical membrane proteins. The procedure was repeated with labeled fragments according to the membrane topology regions. The NMI was found to be 0.34, which indicates almost linear separability of energy profile fragments according to membrane topology regions. The resulting average NMI of 0.14 of globular protein energy profile fragments labeled according to secondary structure elements shows good clustering. This indicates the correlation of protein energy profiles to structural features of proteins and, thus, the correlation to protein structure stability.

2.4. Application of Energy Profiles to V2 Vasopressin Receptor. For investigating the energetics, binding capabilities, and the effect of mutations in the V2 receptor, the structure model of V2R was studied on the level of energy profiles. Therefore, the Molecular Docking Server was used for a docking simulation of the V2R model and the AVP hormone. In [48], it is demonstrated that the semiempirical PM6 partial charges calculation methods, which are implemented in the software of the Molecular Docking server, allowed a docking accuracy of 42 correctly modeled ligand-protein complexes out of a set of 53 ligand-protein complexes determined by X-ray experiments. Regarding the performance of the Molecular Docking Server as well as the energetic trajectories computed while calculating the docking simulation of AVP and V2R (data not shown), the structure of the model and its hormone in bound state was assessed as modeled correctly and used for further analyses. The energy profiles of the V2R model in bound and unbound states were calculated to detect energetic divergences induced by conformational changes during hormone binding.

To detect these binding-induced energetic changes, both derived energy profiles were aligned using a multiple energy

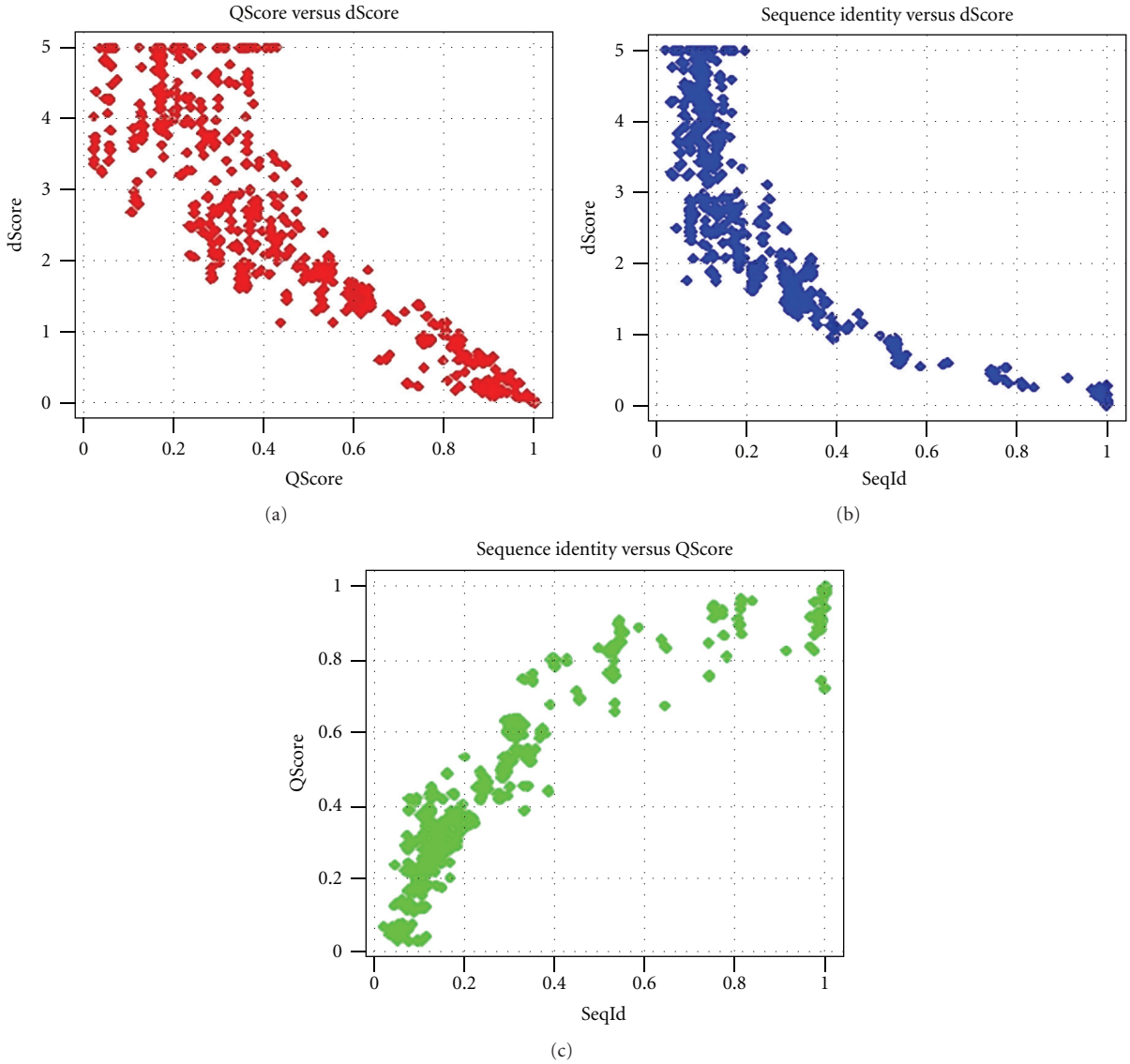


FIGURE 4: Scatter plots of pairwise energy profile distance scores (dScores), structure alignment scores (QScores), and sequence identities of 7 proteins and their homologs. See text for further information.

profile alignment algorithm (MEPAL) which has been adapted from [42]. In the process, all given protein energy profiles are aligned to each other which results in a distance matrix with the corresponding dScores as matrix entries. In the next step, hierarchical clustering is performed using the unweighted pair-group method with arithmetic mean (UPGMA) and the clustering steps, are recorded. Third, according to the tracked clustering steps all energy profiles are introduced in the progressive multiple energy profile alignment using the same techniques as employed in the pairwise energy profile alignment. Thus, significant divergences and similarities in multiple energy profiles can be detected. Furthermore, this method allows the deduction of consensus energy profiles and energy conservation. The energy at each alignment column in the consensus profile is derived by calculating the pairwise energy scores of all energies at this particular position. The energy with the highest sum of these scores is representing the consensus. The conservation is derived by the sum of the pairwise energy scores and is normalized by the theoretically best possible sum. Hence, the optimal conservation in a column equals 1.0 [49].

2.5. Application of Energy Profiles to Aquaporin-2 and Its Homologs. To investigate protein stability in aquaporin-2, a MEPAL of close homolog aquaporins was calculated. The proper proteins are aquaporin-4 (PDB_ID: 3gd8), aquaporin-1 (PDB_ID: 1fqy), the homology model of aquaporin-3 (template PDB_ID: 3ldf), and the structure model of aquaporin-2. The first model was retrieved from the ModBase database and shows less quality than the aquaporin-2 model (data not shown) but is appropriate for further analysis.

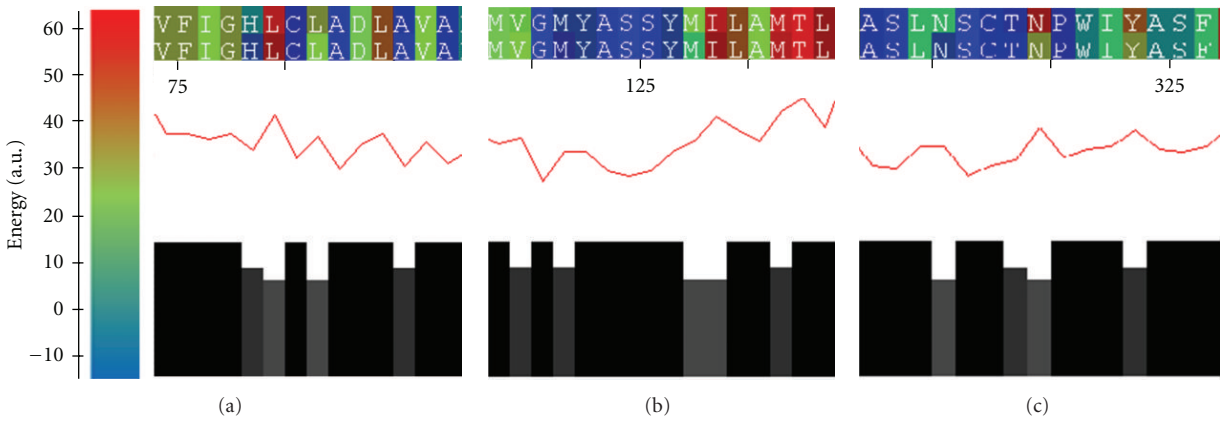


FIGURE 5: The multiple energy profile alignment (MEPAL) of V2R in bound and unbound state to AVP. Although being overall energetically well conserved, three regions showing distinct energy differences can be detected. These regions correspond to residues involved in AVP binding. This observation points to slight changes and rearrangements in the structure of V2R during the binding process.

2.6. Application of Energy Profiles to Aquaporin-2 and Two Well-Described Mutants. For the comparison on the level of energy profiles of aquaporin-2 mutants, two aquaporin-2 models were generated by introducing the two mutations D150E and G196D into the amino acid sequence. Since both mutations are well described in literature [45], correlations of protein energy/stability and experimental observations can be developed.

The modeling of the two mutants was performed by SwissModel [50, 51] using the aquaporin-2 model as template structure. The energy profile of each resulting model was calculated. A MEPAL of the energy profile of the wild-type and the two modeled mutants was generated and investigated for significant differences. Additionally, the distance tree of the energy profiles was computed by means of UPGMA clustering on the basis of the energy profile distance matrix.

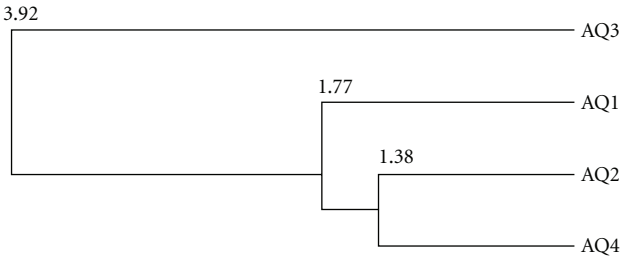


FIGURE 6: The distance tree computed by the energy profile distance matrix of aquaporin-1, -2, -3 and -4 using the unweighted pair group method with arithmetic mean (UPGMA). The energy profiles of aquaporin-1, -2, and -4 show high similarities. The longer distance of aquaporin-3 correlates with higher differences to the energy profiles of the other aquaporins.

3. Results and Discussion

The MEPAL output is separated in three parts. The upper part visualizes the energy profile by coloring the residue one letter codes by their energy. The middle section shows the consensus profile. The conservation is illustrated in the lower part.

The MEPAL of the V2R in bound and unbound state to AVP revealed energetic divergences in the surroundings of the amino acids Ala84 (Figure 5(a)), Ile130 (Figure 5(b)), and Pro322 (Figure 5(c)). This indicates that these amino acids are involved in hormone binding. Their mutations are well described in literature and cause a loss in functionality and hormone affinity [52–56]. This observation emphasizes the quality of the modeled AVP-V2R complex and the coarse-grained energy model discussed in this work. In conclusion, the binding of AVP affects the stability and folding of the structure in only a few spots with rather small structural changes and rearrangements. Hence, this novel energy-profile-based approach brought more evidence and data to the functionality of V2R and the influences of the described mutations.

The analysis of the distance tree generated by the energy profile distance matrix of the investigated aquaporins indicates high similarities between the energy profiles of aquaporin-2 and aquaporin-4. The derived energy profile distance of aquaporin-3 to the other structures corresponds to significant but less similarity (see Figure 6).

The MEPAL alignment of the four aquaporins shows several energetically highly conserved regions. Two of them correspond to the opposite orientated Asn-ProAla motifs. The MEPAL output of the surrounding area of the second Asn-ProAla motif is shown in Figure 7. In this figure, the second Asn-ProAla motif is highlighted by a red box.

The energetic conservation of these motifs and their surrounding amino acids confirms the importance of these residues in water transport. Additionally, the residues Gly188 (highlighted by the right green star in Figure 7), Phe56, Cys189, Ile191, and His180, which are involved in water transport as well, show differences in sequential and energetic conservation (not shown in Figure 7). In more detail, the conserved amino acids Gly188 and Phe56 show

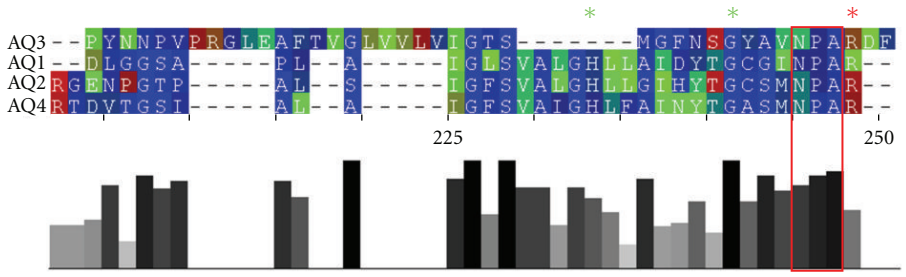


Figure 7: Part of the MEPAL output of aquaporin-1, -2, -3, and -4. Highly sequence conservations are highlighted by red stars. Highlighted by a red box, the second of the two Asn-ProAla motifs can be seen. For detailed discussion, see the text.

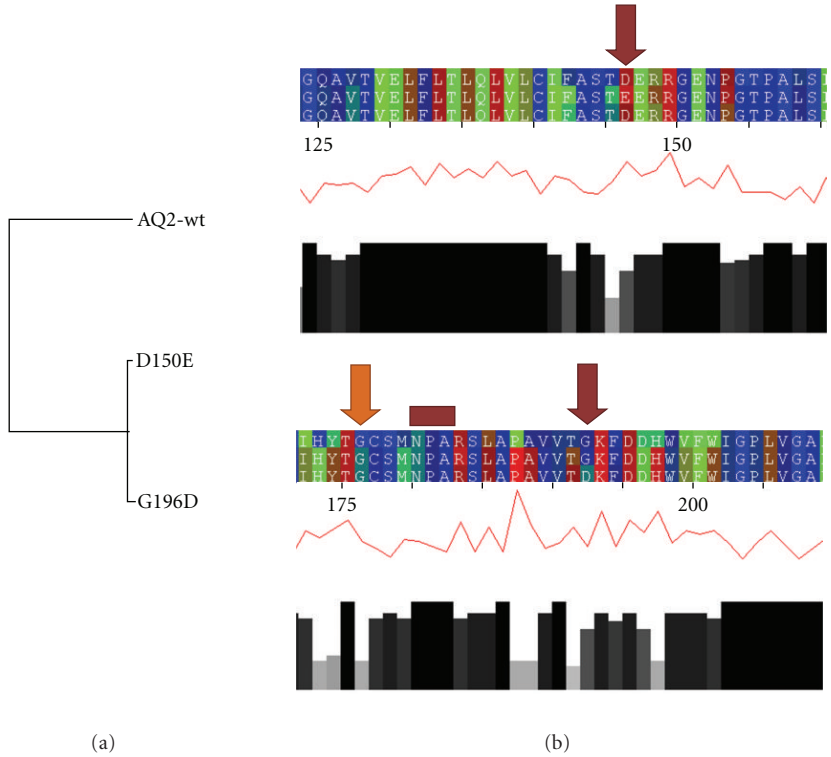


Figure 8: The MEPAL output of the energy profile of the aquaporin-2 wild type and its two mutants D150E and G196D. The distance tree derived by MEPAL indicates the distinct similarity of the energy profiles of the mutants (a). This is visualized by the MEPAL output (b). In the regions, which surround the mutated residues, similarities in energy profile progression can be seen. Thus, it is supported that both mutations affect protein stability and the transcellular water flux as described in literature [45].

slight or no differences at all concerning their energies. Cys189 and Ile191 show no conservation in aquaporin-2, -3, and -4; but these changes have no effect on the level of energy profiles. His180 (highlighted by the left green star in Figure 7) shows sequential and energetic conservation in all aquaporins except aquaporin-3. We postulate that these slight differences do not affect the water flux significantly. A further point of interest lies in Arg195 (highlighted by the red star in Figure 7). This residue is conserved in all four proteins but varies energetically. These differences arise from conformational changes of the residue and the structural environment. Based on the facts we referred to in Section 2.1.2, we postulate that these divergences between aquaporin-1, -2, -3, and -4 lead to a changes in the residue Gly188 and its surrounding residues influencing the transport selectivity and the water flux. It also needs to be said that the significant differences in the energy profile progression between aquaporin-3 and the other structures might result by the less reliable aquaporin-3 model.

The distance tree of the aquaporin-2 wild type and its two modeled mutants (D150E and G196D) indicates strong similarities between the energy profiles of the two mutants (Figure 8(a)). This leads to the conclusion that both mutants induce the same energetic, structural, and functional changes. It needs to be addressed that automated modeling techniques might not be sensitive enough to model

single-point-mutated structures based on a template. Two scenarios arise from these concerns. First, two differing single-point-mutated models and the template differ in structure significantly. Or, as the second scenario, both models and the template show identical folds with respect to backbone conformation and side-chain conformations as well. In the first scenario, the resulting three energy profiles would differ significantly from each other leading to a long-branched, stretched distance tree. In the second scenario, the pairwise comparison of all energy profiles would result to a distance tree with very short branches indicating the given high energy profile similarity. But as seen in the distance tree of the aquaporin-2 wild type and its two modeled mutants (Figure 8(a)), the energy profiles of the mutants match very well and differ to the energy profile of the wild type significantly. This is a strong indication that both models can be assessed as modeled correctly.

While both mutations led to energetic variations in the entire energy profiles, we focused our discussion on the mutation sites (Figure 8(b), red arrows). The mutation D150E induces an energetic increase of the two surrounding residues, thus, decreasing the energetic conservation at these positions (Figure 8(b), top). Interestingly, in this region, the mutation G196D induces almost the same energetic increase as D150E. At the mutation site of the modeled G196D variant, the mutation induces only slight energetic divergences in the sequentially surrounding residues (Figure 4(b), bottom). Furthermore, in this region, the G196D mutation leads to the same energetic variations as the D150E mutation. Interestingly, both mutations do not affect the energetic conservation of the Asn-ProAla motif (highlighted by a red rectangle). Additionally, we point to the energetic changes of G188, a residue involved in water transport (highlighted by an orange arrow). Both mutations lead to an energetic increase of Gly188 and reduce the energetic conservation in these three investigated energy profiles. Thus, it supports the findings that the mutations D150E and G196D affect the transcellular water transport as described in literature [45].

4. Conclusion

We investigated the stability of membrane proteins involved in NDI on the basis of theoretical assumptions. These theoretical methods are based on the so-called energy profile calculation which is demonstrated in this work. On the basis of these stability analyses, we were able to enforce evidence for water flux reduction induced by well-described mutations of aquaporin-2. Furthermore, the correlation of residue and energetic conservation of amino acids involved in water transport was detected. Additionally, we focused on selected point mutations in V2R and their influences in hormone affinity. Based on our data, we were able to enforce evidence described in literature. This indicates that our approach can be successfully employed in the study of other disease-linked membrane protein mutations. Especially, conserved energy profile regions were identified. In general, this approach has proved its applicability regarding similar biological questions.

Acknowledgments

This project was funded by the Free State of Saxony and the University of Applied Sciences Mittweida, Germany. The authors thank Daniel Stockmann for helpful discussions, motivations, and powerful programming.

References

[1] G. E. Tusnády, Z. Dosztányi, and I. Simon, "PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank," *Nucleic Acids Research*, vol. 33, pp. D275–D278, 2005.

[2] G. E. Tusnády, Z. Dosztányi, and I. Simon, "Transmembrane proteins in the Protein Data Bank: identification and classification," *Bioinformatics*, vol. 20, no. 17, pp. 2964–2972, 2004.

[3] P. W. Rose, B. Beran, C. Bi et al., "The RCSB Protein Data Bank: redesigned web site and web services," *Nucleic Acids Research*, vol. 39, supplement 1, pp. D392–D401, 2011.

[4] A. Marsico, D. Labudde, T. Sapra, D. J. Muller, and M. Schroeder, "A novel pattern recognition algorithm to classify membrane protein unfolding pathways with high-throughput single-molecule force spectroscopy," *Bioinformatics*, vol. 23, no. 2, pp. e231–e236, 2007.

[5] A. C. M. Jansen, E. S. van Aalst-Cohen, M. W. Tanck et al., "The contribution of classical risk factors to cardiovascular disease in familial hypercholesterolaemia: data in 2400 patients," *Journal of Internal Medicine*, vol. 256, no. 6, pp. 482–490, 2004.

[6] E. S. van Aalst-Cohen, A. C. M. Jansen, M. W. T. Tanck et al., "Diagnosing familial hypercholesterolaemia: the relevance of genetic testing," *European Heart Journal*, vol. 27, no. 18, pp. 2240–2246, 2006.

[7] M. Sultan, S. Werlin, and N. Venkatasubramani, "The prevalence and characteristics of genetic pancreatitis in children with chronic and recurrent acute pancreatitis," *Journal of Pediatric Gastroenterology and Nutrition*. In press.

[8] G. L. Robertson, "Diabetes insipidus," *Endocrinology and Metabolism Clinics of North America*, vol. 24, no. 3, pp. 549–572, 1995.

[9] S. Ananthakrishnan, "Diabetes insipidus in pregnancy: etiology, eva luation, and management," *Endocrine Practice*, vol. 15, no. 4, pp. 377–382, 2009.

[10] R. Krysiak, I. Kobielusz-Gembala, and B. Okopien, "Recurrent pregnancy-induced diabetes insipidus in a woman with hemochromatosis," *Endocrine Journal*, vol. 57, no. 12, pp. 1023–1028, 2010.

[11] A. M. W. Van Den Ouweland, J. C. F. M. Dreesen, M. Verdijk et al., "Mutations in the vasopressin type 2 receptor gene (*AVPR2*) associated with nephrogenic diabetes insipidus," *Nature Genetics*, vol. 2, no. 2, pp. 99–102, 1992.

[12] W. Rosenthal, A. Seibold, A. Antaramian et al., "Molecular identification of the gene responsible for congenital nephrogenic diabetes insipidus," *Nature*, vol. 359, no. 6392, pp. 233–235, 1992.

[13] P. M. T. Deen, M. A. J. Verdijk, N. V. A. M. Knoers et al., "Requirement of human renal water channel aquaporin-2 for vasopressin-dependent concentration of urine," *Science*, vol. 264, no. 5155, pp. 92–95, 1994.

[14] S. M. Mulders, D. G. Bichet, J. P. L. Rijss et al., "An aquaporin-2 water channel mutant which causes autosomal dominant

nephrogenic diabetes insipidus is retained in the golgi complex," *The Journal of Clinical Investigation*, vol. 102, no. 1, pp. 57–66, 1998.

[15] E. L. Los, P. M. Deen, and J. H. Robben, "Potential of nonpeptide (ant)agonists to rescue vasopressin V2 receptor mutants for the treatment of X-linked nephrogenic diabetes insipidus," *Journal of Neuroendocrinology*, vol. 22, no. 5, pp. 393–399, 2010.

[16] J. H. Robben, N. V. A. M. Knoers, and P. M. T. Deen, "Cell biological aspects of the vasopressin type-2 receptor and aquaporin 2 water channel in nephrogenic diabetes insipidus," *American Journal of Physiology*, vol. 291, no. 2, pp. F257–F270, 2006.

[17] C. Barberis, B. Mouillac, and T. Durroux, "Structural bases of vasopressin/oxytocin receptor function," *Journal of Endocrinology*, vol. 156, no. 2, pp. 223–229, 1998.

[18] M. J. Ślusarz, A. Giełdoń, R. Ślusarz, and J. Ciarkowski, "Analysis of interactions responsible for vasopressin binding to human neurohypophyseal hormone receptors—molecular dynamics study of the activated receptor-vasopressin-Gα systems," *Journal of Peptide Science*, vol. 12, no. 3, pp. 180–189, 2006.

[19] M. J. Ślusarz, E. Sikorska, R. Ślusarz, and J. Ciarkowski, "Molecular docking-based study of vasopressin analogues modified at positions 2 and 3 with N-methylphenylalanine: influence on receptor-bound conformations and interactions with vasopressin and oxytocin receptors," *Journal of Medicinal Chemistry*, vol. 49, no. 8, pp. 2463–2469, 2006.

[20] A. Roy, A. Kucukural, and Y. Zhang, "I-TASSER: a unified platform for automated protein structure and function prediction," *Nature Protocols*, vol. 5, no. 4, pp. 725–738, 2010.

[21] L. Kalé, R. Skeel, M. Bhandarkar et al., "NAMD2: greater scalability for parallel molecular dynamics," *Journal of Computational Physics*, vol. 151, no. 1, pp. 283–312, 1999.

[22] A. D. MacKerell, B. Brooks, and C. L. Brooks, "CHARMM: the energy function and its parameterization with an overview of the program," in *The Encyclopedia of Computational Chemistry. 1*, P. V. R. Schleyer et al., Ed., pp. 271–277, John Wiley & Sons, Chichester, UK, 1998.

[23] L. S. King, D. Kozono, and P. Agre, "From structure to disease: the evolving tale of aquaporin biology," *Nature Reviews Molecular Cell Biology*, vol. 5, no. 9, pp. 687–698, 2004.

[24] T. D. Pollard and W. C. Earnshaw, *Cell Biology*, Springer, Heidelberg, Germany, 2004.

[25] B. L. de Groot, T. Frigato, V. Helms, and H. Grubmüller, "The mechanism of proton exclusion in the aquaporin-1 water channel," *Journal of Molecular Biology*, vol. 333, no. 2, pp. 279–293, 2003.

[26] N. Chakrabarti, E. Tajkhorshid, B. Roux, and R. Pomès, "Molecular basis of proton blockage in aquaporins," *Structure*, vol. 12, no. 1, pp. 65–74, 2004.

[27] N. Chakrabarti, B. Roux, and R. Pomès, "Structural determinants of proton blockage in aquaporins," *Journal of Molecular Biology*, vol. 343, no. 2, pp. 493–510, 2004.

[28] U. Pieper, B. M. Webb, D. T. Barkan et al., "ModBase,a database of annotated comparative protein structure models,and associated resources," *Nucleic Acids Research*, vol. 39, supplement 1, pp. D465–D474, 2011.

[29] L. Willard, A. Ranjan, H. Zhang et al., "VADAR: a web server for quantitative evaluation of protein structure quality," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3316–3319, 2003.

[30] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, no. 4096, pp. 223–230, 1973.

[31] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general Amber force field," *Journal of Computational Chemistry*, vol. 25, no. 9, pp. 1157–1174, 2004.

[32] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Erratum: 'Development and testing of a general amber force field (*Journal of Computational Chemistry* (2004) 25 (1157))'," *Journal of Computational Chemistry*, vol. 26, no. 1, article 114, 2005.

[33] A. D. MacKerell, D. Bashford, M. Bellott et al., "All-atom empirical potential for molecular modeling and dynamics studies of proteins," *Journal of Physical Chemistry B*, vol. 102, no. 18, pp. 3586–3616, 1998.

[34] S. Tanaka and H. A. Scheraga, "Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins," *Macromolecules*, vol. 9, no. 6, pp. 945–950, 1976.

[35] D. H. Wertz and H. A. Scheraga, "Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule," *Macromolecules*, vol. 11, no. 1, pp. 9–15, 1978.

[36] F. Dressel, A. Marsico, A. Tuukkanen, M. Schroeder, and D. Labudde, "Understanding of SMFS barriers by means of energy profiles," in *Proceedings of the German Conference on Bioinformatics*, pp. 90–99, 2007.

[37] J. U. Bowie, R. Luthy, and D. Eisenberg, "A method to identify protein sequences that fold into a known three-dimensional structure," *Science*, vol. 253, no. 5016, pp. 164–170, 1991.

[38] H. Singh and S. Ahmad, "Context dependent reference states of solvent accessibility derived from native protein structures and assessed by predictability analysis," *BMC Structural Biology*, vol. 9, article 25, 2009.

[39] M. J. Sippl, "Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures," *Journal of Computer-Aided Molecular Design*, vol. 7, no. 4, pp. 473–501, 1993.

[40] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[41] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981.

[42] D. G. Higgins, J. D. Thompson, and T. J. Gibson, "[22] Using CLUSTAL for multiple sequence alignments," *Methods in Enzymology*, vol. 266, pp. 383–400, 1996.

[43] K. A. Dill and H. S. Chan, "From levinthal to pathways to funnels," *Nature Structural Biology*, vol. 4, no. 1, pp. 10–19, 1997.

[44] E. Krissinel and K. Henrick, "Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions," *Acta Crystallographica Section D*, vol. 60, no. 12, part 1, pp. 2256–2268, 2004.

[45] C. Guyon, Y. Lussier, P. Bissonnette et al., "Characterization of D150E and G196D aquaporin-2 mutations responsible for nephrogenic diabetes insipidus: importance of a mild phenotype," *American Journal of Physiology*, vol. 297, no. 2, pp. F489–F498, 2009.

[46] T. M. Martinetz and K. J. Schulten, "A neural-gas network learns topologies," in *Artificial Neural Networks*, T. Kohonen, K. Mkisara, O. Simula, and J. Kangas, Eds., pp. 397–402, North-Holland, Amsterdam, The Netherlands, 1991.

[47] I. H. Witten and F. Eibe, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Amsterdam, The Netherlands, 2005.

[48] Z. Bikadi and E. Hazai, "Application of the PM6 semi-empirical method to modeling proteins enhances docking accuracy of AutoDock," *Journal of Cheminformatics*, vol. 1, no. 1, article 15, 2009.

[49] F. Heinke, A. Tuukkanen, and D. Labudde, "Analysis of Membrane Protein Stability in Diabetes Insipidus," K. Kamoi, Ed., InTech, 2011.

[50] K. Arnold, L. Bordoli, J. Kopp, and T. Schwede, "The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling," *Bioinformatics*, vol. 22, no. 2, pp. 195–201, 2006.

[51] F. Kiefer, K. Arnold, M. Künzli, L. Bordoli, and T. Schwede, "The SWISS-MODEL Repository and associated resources," *Nucleic Acids Research*, vol. 37, no. 1, pp. D387–D392, 2009.

[52] E. Albertazzi, D. Zanchetta, P. Barbier et al., "Nephrogenic diabetes insipidus: functional analysis of new AVPR2 mutations identified in Italian families," *Journal of the American Society of Nephrology*, vol. 11, no. 6, pp. 1033–1043, 2000.

[53] K. Pasel, A. Schulz, K. Timmermann et al., "Functional characterization of the molecular defects causing nephrogenic diabetes insipidus in eight families," *Journal of Clinical Endocrinology and Metabolism*, vol. 85, no. 4, pp. 1703–1710, 2000.

[54] J. H. Robben, N. V. A. M. Knoers, and P. M. T. Deen, "Characterization of vasopressin V2 receptor mutants in nephrogenic diabetes insipidus in a polarized cell model," *American Journal of Physiology*, vol. 289, no. 2, pp. F265–F272, 2005.

[55] D. Morin, Y. Ala, N. Sabatier et al., "Functional study of two V2 vasopressin mutant receptors related to NDI P322S and P322H," *Advances in Experimental Medicine and Biology*, vol. 449, pp. 391–393, 1998.

[56] R. Vargas-Poussou, L. Forestier, M. D. Dautzenberg, P. Niaudet, M. Déchaux, and C. Antignac, "Mutations in the vasopressin V2 receptor and aquaporin-2 genes in 12 families with congenital nephrogenic diabetes insipidus," *Journal of the American Society of Nephrology*, vol. 8, no. 12, pp. 1855–1862, 1997.

2.2 Conference Presentation - Predicting functionality of the non-expressed putative human OHCU decarboxylase by means of novel energy profile-based methods

This scientific contribution was part of the 13. Nachwuchswissenschaftlerkonferenz 2012. It included a 20 minute talk as well as a publication in the conference proceedings. This work focuses on the putative, non-expressed human OHCU decarboxylase. OHCU decarboxylase is an essential enzyme in uric acid degradation. The missing of OHCU decarboxylase is linked to diseases caused by high uric acid concentration in the human blood, i.e. renal calculus formation, hyperuricaemia and gout. However, the causes for the missing expression have not been clarified to date. Major causes might be disabled transcription mechanisms or the loss of the gene's protein-coding ability. In this work, the energy profile of a structure model of the putative human OHCU decarboxylase is computed and compared to the energy profile predicted from sequence. In the theory sections the energy profile prediction algorithm (eGOR) is discussed in detail. Finally, energy profile data computed from known OHCU decarboxylase structures and data predicted from sequence are arranged and compared to the energy profiles of human OHCU decarboxylase. Based on the derived results it can be postulated that the missing expression is rather caused by dysfunctional transcription and translation mechanisms than the lack of enzymatic activity as a result of protein-coding inability.

Predicting functionality of the non-expressed putative human OHCU decarboxylase by means of novel protein energy profile-based methods

Florian Heinke*, B. Sc., E-Mail: florian.heinke@hs-mittweida.de, Hochschule Mittweida, University of Applied Sciences, Technikumplatz 17, D-09648 Mittweida

Prof. Dr. rer. nat. Dirk Labudde*, E-Mail: dirk.labudde@hs-mittweida.de, Hochschule Mittweida, University of Applied Sciences, Technikumplatz 17, D-09648 Mittweida

Abstract

Renal calculus, hyperuricaemia, gout and Lesch-Nyhan syndrome are diseases that originate from disorders in the human purine catabolism. Both former diseases can be a result of diabetes mellitus or long-term, one-sided, purine-rich feeding and alcohol consumption. The latter is an X-linked recessive inheritable disease with low incidences in approximately one per 380,000 live births. Disorders in the human purine catabolism lead to a 15-25-fold increase in uric acid concentration in the blood (hyperuricaemia). Due to its low water solubility, uric acid crystals can develop and can lead to renal calculus or gout.

Since purine is a main component of DNA and RNA molecules, purine degradation is a catabolism present in the majority of all organisms. Commonly, as the last steps in purine catabolism, uric acid is degraded by three enzymatic reactions: it is catalysed by urate oxidase, HIU hydrolase and OHCU decarboxylase, respectively. Despite being present in the human chromosomal genome, none of the three enzymes are expressed, which leaves uric acid as the final degradation product. Such non-expressed genes are called “pseudo genes” and might result by gene duplication and/or deactivation by genomic mutations.

In this work, novel protein energy profile methods are employed for predicting the activity of the putative human OHCU decarboxylase encoded by the OHCU pseudo gene. The energy profile of a modelled human OHCU decarboxylase structure is compared to its predicted energy profile. This data is arranged with energy profiles derived from experimental structure and sequence data of homolog OHCU decarboxylases of various organisms. Based on our data, we postulate that the putative human OHCU decarboxylase shows OHCU decarboxylation activity; its missing expression is rather caused by missing transcription mechanisms than the lack of enzymatic activity.

Keywords: purine catabolism, putative human OHCU decarboxylase, energy profile calculation, energy profile prediction

1 Introduction

Pseudo genes are usually the result of gene duplication. In most cases, one of the duplicates loses functional activity due to genomic mutations during evolution, and its expression is down regulated while the other duplicate is still expressed [1]. In this work, the putative human 2-oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline (OHCU) decarboxylase is investigated. This 173 amino acid enzyme is encoded by the PRHOXNB pseudo gene located at locus 13q12.2. The existence of this gene was predicted using gene prediction algorithms, but to this day, the expression of this protein cannot be detected by gene expression analysis [3]. In contrast to common pseudo genes, no duplicate could be located in the human genome, which poses the question of whether this gene was silenced due to dysfunctional mutations or due to a decrease of transcriptional factors.

OHCU decarboxylase plays a major role in purine catabolism in most organisms. Here, in the final steps of purine degradation, uric acid is oxidated to 5-

* Author to whom correspondence should be addressed

hydroxyisourate (HIU) by urate oxidase. HIU is converted to OHCU by HIU hydrolase, and is finally decarboxylated to (S)-allantoin by OHCU decarboxylase. An experimentally derived OHCU structure bound to (S)-allantoin is illustrated in Figure 1. Those last three enzymatic steps are missing in human purine catabolism, leaving uric acid as the final degradation product, which, due its low water solubility, is a main factor in forming diseases like renal calculus and hyperuricaemia. As another result, the uric acid concentration in human blood is about tenfold higher than in primates, birds and reptiles. One advantage of this high uric acid concentration is its property as an antioxidant; it serves as a defense mechanism against DNA-damaging toxins [2,3].

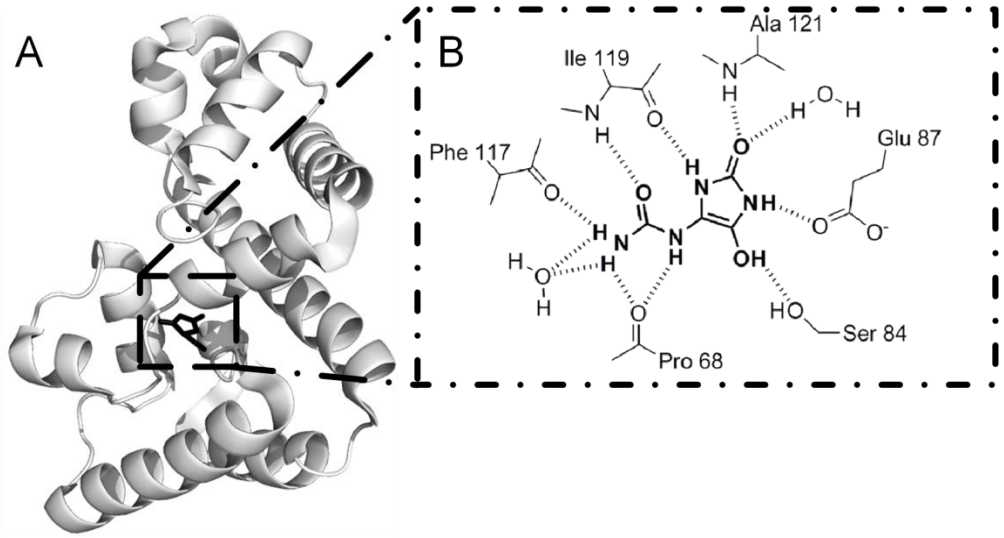


Figure 1: **(S)-allantoin binding of a homolog OHCU decarboxylase.** A: Structure of OHCU decarboxylase of the zebrafish, bound to (S)-allantoin (black) (PDB_ID 2o73), which is buried deep inside the structure. Hence, structural rearrangements occur during substrate binding. B: The schematic network of amino acid residues involved in binding. Residues are labeled by name and sequence index [4].

In this work, the energy profile of a protein structure model of the putative human OHCU decarboxylase is generated and compared to the energy profile which was predicted by the human OHCU decarboxylase amino acid sequence. This data is arranged with energy data derived from experimental structure and sequence data. Based on the results from these analyses, the human OHCU decarboxylase is still in a functional state, but is not expressed due to probably missing transcriptional mechanisms.

2 Materials and methods

2.1 Theory of energy profiles

The coarse-grained energy model presented here can be assigned to the so-called knowledge-based potentials approaches, where Boltzmann principles and rules of statistical physics are employed. The correlations to protein structure and protein structural/functional features, as well as the application of energy profiles, are demonstrated in [5,6]. In this work, the basic principles are summarised.

In this model, the structural environment of an observed residue (equivalent to an amino acid in a protein structure) is transformed into a one-dimensional representation by deriving its pseudo-potential in this particular environment. The pseudo-potentials are based on the tendency of each of the 20 canonical amino acids either being buried or exposed to the surrounding solvent. Here, the Boltzmann principle is employed for these tendencies, leading to the approximation of the pseudo-potential of the observed residue. To derive these tendencies, an inside/outside property was implemented [5]. In this, *i* denotes one of the 20 canonical amino acids; $n_{in,i}$ and $n_{out,i}$ correspond to the absolute frequencies of *i* being assigned as inside or outside by the inside/outside property, respectively. The inside/outside property is denoted as

$$f(i) = \begin{cases} n_{\mathrm{in},i}{+}{+}, & |C_{\alpha,i} - c| < 5\text{Å} \vee (C_{\alpha,i} - C_{\beta,i})(C_{\alpha,i} - c) < 0 \\ n_{\mathrm{out},i}{+}{+}, & \text{else} \end{cases} \tag{1}$$

where *c* defines the center of mass of all C_α atoms in a sphere of r = 5 Å surrounding *i.* Using the inverse Boltzmann principle, the pseudo-energy e^*_i can be approximated.

$$e_i^* = -\ln\left(\frac{n_{\mathrm{in},i}}{n_{\mathrm{out},i}}\right) \tag{2}$$

Furthermore, the pairwise interaction energy of *i* corresponds to its environment inside the structure. Here, the expected tendency of the environment can be computed and used for pseudo-energy approximation:

$$\begin{aligned} P_{k\in\mathrm{Env}} &= \prod_{k\in\mathrm{Env}} p_k = \prod_{k\in\mathrm{Env}} \left(\frac{n_{\mathrm{in},k}}{n_{\mathrm{out},k}}\right) \\ \ln P_{k\in\mathrm{Env}} &= \sum_{k\in\mathrm{Env}} \ln\left(\frac{n_{\mathrm{in},k}}{n_{\mathrm{out},k}}\right) \end{aligned} \tag{3}$$

The environment was defined by the contact function *g(i,j),* which states

$$g(i,j) = \begin{cases} 1, & \|C_{\alpha,i} - C_{\alpha,j}\| < 8\text{Å} \\ 0, & \text{else} \end{cases} \tag{4}$$

Applying the inverse Boltzmann principle to the expression given in (3) with respect to the environment defined by (4), the total energy of *i* equals

$$E_i^* = \sum_{j\in S\setminus i} \left[g(i,j) \left(\ln\left(\frac{n_{\mathrm{in},j}}{n_{\mathrm{out},j}}\right) + e_i^* \right) \right] \tag{5}$$

where *S* corresponds to a given protein structure. In this work, the inside/outside statistics were derived by employing the inside/outside property to a set of 2700 non-redundant globular proteins. The energy profile of a protein corresponds to the n-tuple of all E_i^*. Hence, an energy profile represents physico-chemical protein properties as a one-dimensional vector.

In [5,6], it is shown that energy profiles can be compared by means of dynamic programming consensus string optimisation using optimisation weights based on Gaussian integral intervals of the energy distribution. Additionally, energy profile distances can be computed and used for hierarchical clustering [6].

2.2 eGOR2 - Predicting discretised energy profiles by amino acid sequence

The novel eGOR2 algorithm performs an energy profile prediction using the amino acid sequence of a protein as input. Its theory stems from the protein secondary structure prediction algorithm GOR-III [1].

To ensure the predictability of usually continuous energy values, the energy distribution was discretised using quartiles, reducing the prediction and training alphabet to the size of 4. Thus, each energy value in a training set can be assigned to its quartile and conditional probabilities, and conditional frequencies can be computed and used as input for training the eGOR2 statistics. The statistics are derived by counting the absolute frequencies of each canonical amino acid *i* in each quartile as well as the frequencies of *i* occurring in a given quartile with respect to the residue composition of a sequence frame of 7 residues surrounding *i*. From these statistics, the expected information difference of each quartile can be approximated by a given amino acid sequence, and the corresponding discretised energy profile can be computed. Formally, the expected information difference of a frame of 7 residues is predicted by the following approximation:

$$I\left(\Delta Q_i^q; R_1 \cap \ldots \cap R_n\right) \approx \log\left[\frac{f_{Q^q,R}}{f_{n-Q^q,R}}\right] + \sum_{m=-3,m\neq 0}^{+3} \log\left[\frac{f_{Q_i^q,R_{i+m},R_i}}{f_{n-Q_i^q,R_{i+m},R_i}}\right] + \log\left[\frac{f_{n-Q_i^q,R_i}}{f_{Q_i^q,R_i}}\right] \quad (6)$$

Here Q_i^q corresponds to the quartile in which the information difference is predicted regarding to the observed residue R_i. The total number of observed residues in the training set is denoted by *n*. All conditional frequencies that are needed for quartile prediction can be read from the training set. The predicted quartile corresponds to the quartile with the highest information difference.

$$Q_i^q = \operatorname{argmax}\begin{cases} I\left(\Delta Q_i^1; R_{i-3} \cap \ldots \cap R_{i+3}\right) \\ \ldots \\ I\left(\Delta Q_i^4; R_{i-3} \cap \ldots \cap R_{i+3}\right) \end{cases} \quad (7)$$

To evaluate the eGOR2 algorithm, ten-fold cross-validation was performed. Each cross-validation run included the retraining of the eGOR2 statistics by the randomly chosen training set and the application of these statistics in the energy profile predictions of the validation set. The ten-fold cross-validation shows an average identity rate of 0.67 with σ = 0.05 and a similarity rate of 0.98 with σ = 0.015, as well as a progression prediction accuracy of 0.75 with σ = 0.05. Identities correspond to correctly predicted quartiles ($q_{pred} = q_{real}$). A prediction is considered similar if $|q_{pred} - q_{real}| = 1$. This prediction accuracy is sufficient for molecular biological analyses.

2.3 Calculation and prediction of OHCU decarboxylase energy profiles

First, a structure model of human OHCU was found at the ModBase database (see Figure 2A). ModBase is a database of protein structure models [1]. These models are generated automatically and can be updated by users on demand. The human OHCU model was evaluated by the VADAR server, where modelling reliability is validated by checking the structure for stereo-chemical misfoldings. The model showed a high modelling reliability indicated by high 3D quality indices (see Figure 2B), and is thus sufficient for further analyses. After calculating the energy profile based on the structure (see Figure 2C, grey), the energy profile of the human OHCU decarboxylase was predicted by sequence using eGOR2 (see Figure 2C, black).

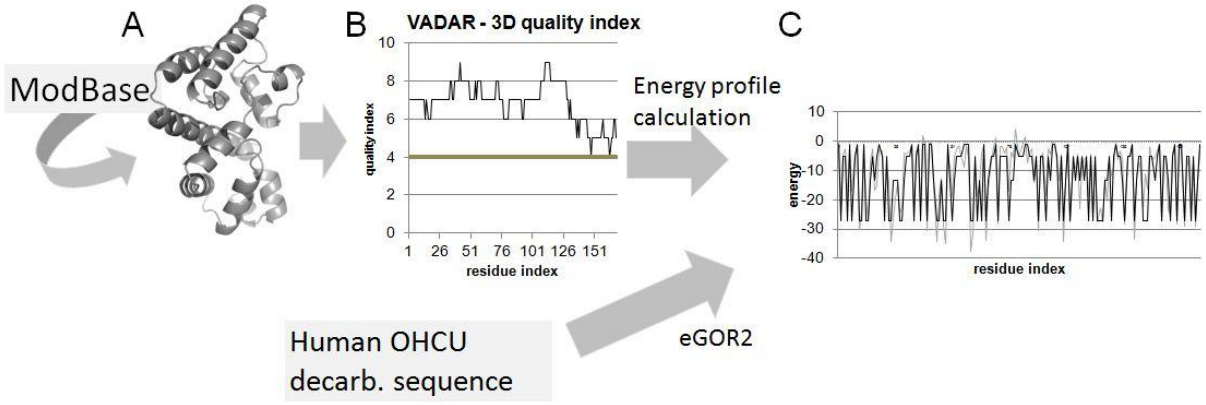


Figure 2: **Workflow of initial human OHCU decarboxylase analysis.** A: Structure model of human OHCU decarboxylase was downloaded from ModBase, which stores automatically generated protein structure models. The model reliability was validated using the VADAR server (B), which indicated high reliability (quality indices greater than 4). This model was used for energy profile calculation and was compared to the energy profile predicted by eGOR2 based on the human OHCU sequence (C).

To include evolutionary data, family seed sequences of the OHCU decarboxylase family were downloaded from the Pfam database (Pfam-Id: PF09349) and were reduced further using sequence clustering at a threshold of 50 % sequence identity. The resulting 53 sequences were used as input for eGOR2 energy profile predictions. Additionally, the energy profile of the known OHCU decarboxylase structure from zebrafish (PDB_ID: 2o70) and *Arabidopsis thaliana* (PDB_ID: 2q37) were included. Other known structures were omitted due to high sequence and high energy profile redundancy. To reveal similar energy profiles, hierarchical neighbor-joining clustering applied to all pairwise energy profile distances was computed. Afterwards, detected similar energy profiles were compared.

3 Results and discussion

The hierarchical clustering of all energy profile distances revealed similarities between human OHCU decarboxylase and OHCU decarboxylase of *Arthrobacter aurescens* (A1RAF3) (data not shown). The comparison of similar energy profiles revealed energetic conservations of all residues involved in substrate binding (see Figure 1B) and at the catalytic active His67. Figure 3 illustrates these correspondences.

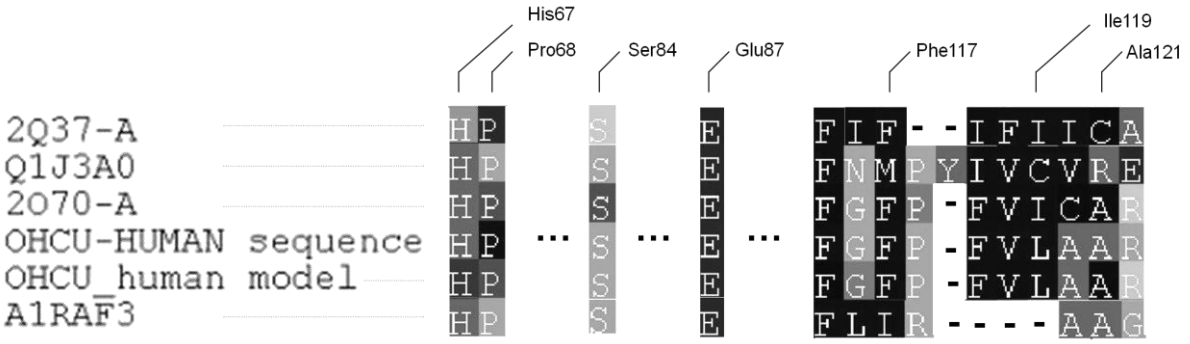


Figure 3: **Manual alignment of OHCU decarboxylase energy profiles.** The predicted energy profile of the human OHCU decarboxylase corresponds well to the energy profile calculated by the structure model. Here, the residues involved in catalysis and substrate binding are illustrated and coloured according to their energy (black - low energy, light grey - high energy). Similar predicted energy profiles and energy profiles calculated on structural data match well, which leads to the conclusion, that the putative human OHCU decarboxylase is still functional if expressed.

4 Conclusion

In this work, the functionality of putative human OHCU decarboxylase was predicted theoretically by means of novel energy profile-based methods. A structure model of this enzyme was validated and used for energy profile calculation. Based on the enzymes sequence, an energy profile was predicted by applying the eGOR2 algorithm. Energetic similarities were detected in the OHCU decarboxylase family. Since these similarities match the predicted energy profile of human OHCU decarboxylase and the energy profile derived by the structure model, the functionality of putative human OHCU decarboxylase is predicted as still active. We postulate that the silenced expression of the pseudo gene leads back rather to missing transcriptional mechanisms than to dysfunctional mutations.

References

[1] Zvelebil MJ, Baum JO. Understanding bioinformatics. 1st ed. New York: Garland Science; 2008.

[2] Rehm H, Hammar F. Biochemie light. 3rd ed. Frankfurt a.M.: Harri Deutsch; 2005

[3] Ramazzina et al. The structure of OHCU decarboxylase provides insights into the mechanism of uric acid degradation. J Biol Chem. 2007 Jun 22;282(25):18182-9.

[4] French JB, Ealick SE. Structural and mechanistic studies on *Klebsiella pneumoniae* OHCU decarboxylase. J BiolChem. 2010 Nov 12;285(46):35446-54.

[5] Heinke F, Brumm R. Energieprofilbasierende Analysemethoden von Proteinfamilien. In: Stolzenburg F, Ruh F, editors. Proceedings of the NWK 12; 2010, Wenigerorde: Germany

[6] Heinke F, Tuukkanen A, Labudde D (2011). Analysis of Membrane Protein Stability in Diabetes Insipidus, Diabetes Insipidus, Kyuzi Kamoi (Ed.), ISBN: 978-953-307-367-5, InTech

2.3 Conference Posters and Poster Abstracts

2.3.1 Energy profile based protein structure description - comparison and description

This poster abstract and poster presentation contributed to the German Conference on Bioinformatics 2010. The potentials of energy profile-based analyses are emphasized. Additionally, the eProS database and toolbox are included in the discussions.

Energy based Prediction and Comparison of Protein Structures

Florian Heinke, Stefan Schildbach and Dirk Labudde

University of Applied Sciences Mittweida
fheinke@hs-mittweida.de

A lot of tools and methods in the field of bioinformatics and structure biology are based on structure and/ or sequence comparison. In our work we demonstrate a new method based on so called energy profiles for comparison and prediction of globular protein structures. Those profiles are calculated by coarse grained models [1,2]. Based on the residue contacts in known protein structures, we calculated the potential for pair wise residue-residue-interactions [2]. An energy profile is a schematic plot of the interaction energy of each residue as a function of the residue position in the sequence. As an abstraction of protein sequence and structure information, each energy profile represents a protein as a fingerprint.

In our new approach, we perform an energy profile alignment based on the Needleman-Wunsch-Algorithm by using self constructed cost-schemes [3]. These cost schemes are derived by the statistical energy distribution of more than 4000 known structures. We showed that the z-scores of these pair wise energy profile alignments correlate with structural alignment scores performed by known methods.

Additionally, we developed a modified GOR Algorithm for the prediction of energy profiles. By using the energy profile alignment algorithm, this prediction leads to a hit list of similar structures [3]. These described tools can be accessed at http://bioservices.hs-mittweida.de/Epros/.

References

[1] Wertz, D. H. and Scheraga, H. A. (1978). Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule. Macromolecules, 11(1), 9–15.

[2] F.Dressel, A. Marsico, A. Tuukkanen, M. Schroeder and Dirk Labudde; Understanding of SMFS barriers by means of energy profiles; Proc. GCB (2007)

[3] Bachelor Thesis: Florian Heinke, Hochschule Mittweida, 2010, Energieprofilbasierende Methoden zur Analyse von Proteinfamilien.

Energy profile based protein structure description - comparison and prediction -

Florian Heinke, Steffan Schildbach and Dirk Labudde
University of Applied Sciences, Technikumplatz 17, D-09648 Mittweida

HOCHSCHULE MITTWEIDA UNIVERSITY OF APPLIED SCIENCES

Background and Motivation

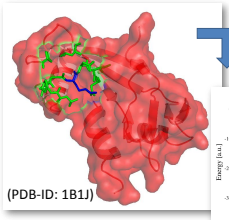
Proteins are important bio-molecules in biological systems and activities. Knowledge of the protein structure gives us insight into function of the protein and its dynamics. On the other side protein structure comparison is a fundamental task in structure biology. The number of protein structures has grown rapidly over the last decade. There is a need for new techniques which can rapidly compare protein structures with high performance and accuracy. Protein structure comparison is crucial for understanding protein evolution, architecture and function.
A lot of tools and methods in the field of bioinformatics and structure biology are based on structure and/or sequence Comparison [1,2,3]. **In this work we demonstrate a new method based on so called energy profiles for comparison and prediction of globular protein structures.**

Methods

Theory of energy profiles

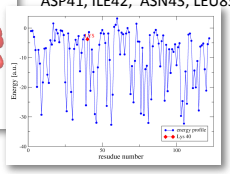
- define inside and outside residues in protein structures
- statistics - count the inside and outside residues
- define an energy per residue contact

$$e_{aa} \propto \ln\left(\frac{n_{aa}^{in}}{n_{aa}^{out}}\right) \longrightarrow e_{aa_{ij}} = (e_{aa_i} + e_{aa_j}) \longrightarrow E_{aa_i} = \frac{1}{2}\sum_{<i,j>} e_{aa_{i,j}}$$

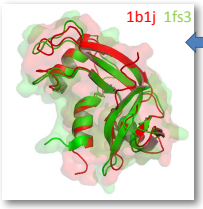


sum over the pair wise potential of each residue

Lys 40 (LEU35, THR36, PRO38. CYS39, ASP41, ILE42, ASN43, LEU83, PRO88)

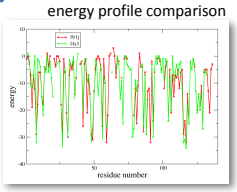


Alignment of energy profiles - *eAlign*



- dynamical programming
- Needleman-Wunsch algorithm

cost sheme [4] optimized for energy profile comparison



Sequence identity: 30,3%
Sequence similarity: 47,0%

DaliLite zScore: 15,3
eAlign zScore: 1,9

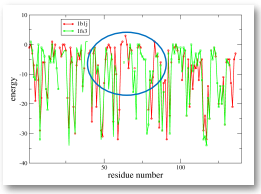
```
1b1j: DNSRYTHFLTQATDAKPQ--GRDDRYCKSIMRRRGLY--SPCKDINTFIRGNERSIKAICEN-KN
1fs3: -ETAAAKFERQHMDSSTSAA-SSSNYCNQMMKSRNLTKD-RCKPVNTFVHESLADVQAVC--S--

GMP-R-KEN----LRISKSSPQVTTCKLH-GGSP-WPPCQYRATAGFRWVVVACE-NG-LPVHLDQSIFRR
QKNVACKNQQTNCYQSY-STMSIT--DCRETGSSKYPNCAYKTTQANKHIIVACEGNPYVPVHFDAS----
```

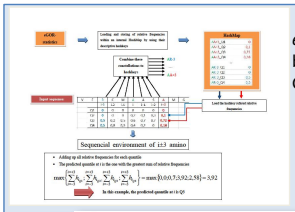
Results and Application

The analysing of the alignment leads to the conclusion, that certain parts of the alignment are significantly more conserved in both proteins, than other regions. In the moment, our routine contains tools for energy profile **calculation, comparison** and **prediction**.

1b1j 1fs3

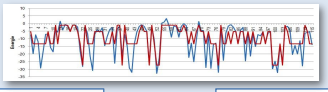


Prediction of energy profiles - *eGOR*

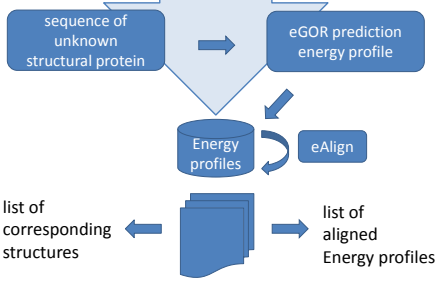


eGOR algorithm based on the classical GOR algorithm [5]

The calculated e-profile from the structure (blue) and the predicted profile (red) - (PDB-ID: 1B1J)



Searching of corresponding structures - *eSearch*



Server : *ePros*: bioservices.hs-mittweida.de/Epros [6]

References

[1] Holm L and Sander C: Mapping the protein universe. Science (1996, 273 (5275) :595–603,
[2] Orengo CA and Taylor WR: SSAP: sequential structure alignment program for protein structure comparison. Methods in Enzymology 1996, 266:617–635.
[3] Konagurthu AS, Whisstock JC, Stuckey PJ and Lesk AM: MUSTANG: A multiple structural alignment algorithm. Proteins: Structure, Function and Bioinformatics 2006, 64(3):559–574.
[4] Bachelorarbeit: Florian Heinke, Hochschule Mittweida, 2010, *Energieprofilbasierende Methoden zur Analyse von Proteinfamilien.*
[5] J. Garnier, D.J. Osguthorpe, and B. Robson (1978) Predicting the secondary structure of globular proteins , *J. Mol. Biol.* 120: 97-120
[6] Bachelorarbeit: Steffan Schildbach, Hochschule Mittweida, 2010, *Energieprofilbasierende Suchen und Structurvergleiche*

contact: dirk.labudde@hs-mittweida.de

2.3.2 Novel prediction algorithm eGOR - from sequence to stable regions of membrane proteins

This conference contribution introduces the eGOR algorithm as an approach for predicting stable features and so-called unfolding barriers predictable by SMFS in membrane proteins.

Novel prediction algorithm eGOR - from sequence to stable regions of membrane proteins

Riccardo Brumm, Eric Frenzel and Florian Heinke

Hochschule Mittweida, University of Applied Sciences
Technikumplatz 17, D-09648 Mittweida, Germany
email: forename.surname@hs-mittweida.de

Abstract

Membrane proteins are important elements in signal transduction and transfer of agents between cells. Often the occurrence of point mutations is sufficient to influence protein stability or functionality [1]. Thus, detecting stable regions in protein structures plays an important role in understanding protein functionality. Membrane proteins pose an experimental challenge and lead to limited data and knowledge. Therefore deriving data by other non-experimental methods is indispensable. The claim of this work is to detect stable regions in membrane protein structures by a new prediction algorithm.

We calculated amino acid interaction potentials based on residue-residue contacts in known membrane protein structures. Using these potentials an energy profile of a membrane protein based on its structural information can be derived. An energy profile is a schematic plot of the interaction energy of each residue as a function of the residue position in the sequence. In conclusion, stable regions of membrane proteins can be predicted. These predictions were evaluated against MD simulation and single molecular force spectroscopy (SMFS). We showed that stable regions detected by SMFS correlate with stable regions predicted by energy profiles [2]. Furthermore, we developed a GOR-based prediction algorithm, which performs an energy profile prediction based on an amino acid sequence - eGOR. The predicted energy profiles can be applied to our approach. In conclusion, unfolding barriers and structural information can be derived eGOR [3].

The eGOR prediction algorithm and some energy profile based tools are available at http://bioservices.hs-mittweida.de/Epros/.

References

[1] Riccardo Brumm. Energieprofilbasierende stabilitätsanalyse von membranproteinen. Master's thesis, University of Applied Sciences Mittweida, 2010.

[2] Dressel F., Marsico A., Tuukkanen A., Schroeder M., and Labudde D. Understanding of SMFS barriers by means of energy profiles. In *Proc GCB 2007*, 2007.

[3] Florian Heinke. Energieprofilbasierende methoden zur analyse von proteinfamilien. Master's thesis, University of Applied Sciences Mittweida, 2010.

Novel prediction algorithm eGOR – from sequence to stable regions of membrane proteins

Riccardo Brumm, Florian Heinke, Eric Frenzel and Dirk Labudde
University of Applied Sciences, Technikumplatz 17, D-09648 Mittweida

HOCHSCHULE MITTWEIDA UNIVERSITY OF APPLIED SCIENCES

Background and Motivation

Membrane proteins are important elements in signal transduction and transfer of agents between cells. Often the occurrence of point mutations is sufficient to influence protein stability or functionality [1]. Thus, detecting stable regions in protein structures play an important role in understanding protein functionality. Membrane proteins pose an experimental challenge which leads to limited data and knowledge. Therefore deriving data by other non-experimental methods is indispensable. The claim of this work is to detect stable regions in membrane proteins by a new prediction algorithm.

Methods

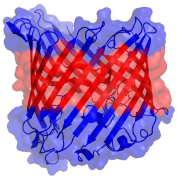
Theory of energy profiles

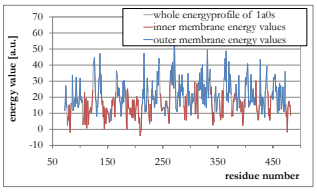
- define inside and outside residues in protein structures
- statistics - count the inside and outside residues and distinguish between inner and outer membrane amino acids
- calculation of amino acid interaction potentials based on residue-residue contacts

$$e'_{i0} = -k_B T \ln\left(\frac{n_{in}}{n_{out}}\right) \Rightarrow e_{i0} = \frac{1}{\alpha_i} e'_{i0} \Rightarrow e_{ij} = e_{i0} + e_{j0} + e^*_{ij}$$

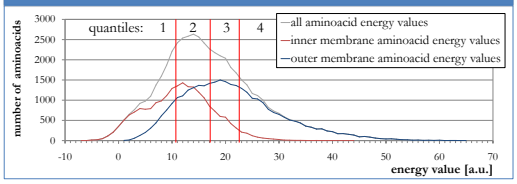
$$e^*_{ij} = -k_B T \ln\left(\frac{n_{ij}}{N_{contact} P_i P_j}\right) \Rightarrow e_{ij} = e_{i0} + e_{j0} + e^*_{ij}$$

Using these potentials an energy profile of a membrane protein based on its structural information and transmembrane regions can be derived.



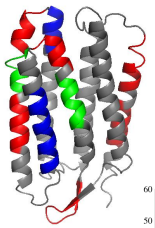


An energy profile is a schematic plot of the interaction energy of each residue as a function of the residue position in the sequence. Therefore stable residues (quantile 1) and unstable residues (quantile 4) can be predicted. Quantiles 2 and 3 are ambivalent.



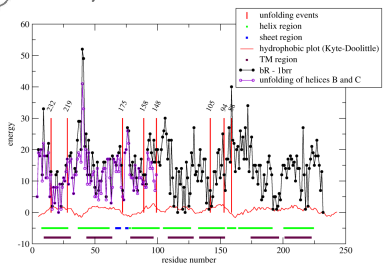
Results and Application

Energy profile prediction evaluation against SMFS



The structural barriers of **1brr** are categorized:

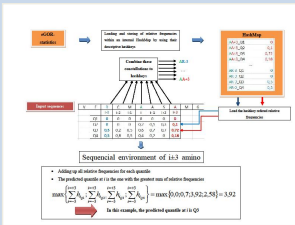
- SMFS measurement confirmed by energy profile calculation is coloured **red**
- detected by SMFS measurement but not by energy profile calculation is coloured **green**
- detected by energy profile calculation but not by SMFS is coloured **blue**

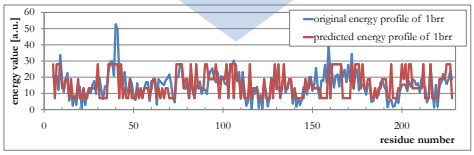


eGOR-based prediction algorithm

eGOR algorithm based on the classical GOR algorithm [5]

The calculated e-profile from the structure (**blue**) and the predicted profile (**red**) - (PDB-ID: 1brr)





In conclusion unfolding barriers and structural information can be derived by *eGOR*.

References

[1] Bachelor Thesis: Riccardo Brumm. Energieprofilbasierende Stabilitätsanalyse von Membranproteinen. Fakultät Mathematik/Naturwissenschaften/Informatik, Hochschule Mittweida, University of Applied Sciences : 2010.
[2] F. Dressel, A. Marsico, A. Tuukkanen, M. Schroeder, D. Labudde. Understanding of SMFS barriers by means of energy profiles. Proc GCB 2007.
[3] Bachelor Thesis: Florian Heinke. Energieprofilbasierende Methoden zur Analyse von Proteinfamilien. Fakultät Mathematik/Naturwissenschaften/Informatik, Hochschule Mittweida, University of Applied Sciences : 2010.
[4] ***ePros-Server.*** bioservices.hs-mittweida.de/Epros

contact: riccardo.brumm@hs-mittweida.de

2.3.3 The novel approach eHMM for analyzing membrane proteins in case of HP_0565

HP_0565 is a membrane protein present in *Helicobacter pylori* that is hypothesized to be triggering pathogenicity. In this work, an approach for predicting membrane spanning regions in polytopic α-helical membrane proteins based on predicted energy profiles (eHMM). Based on meta-predictions obtained from various methods and the eHMM prediction, unfolding barriers in HP_0565 detected by SMSF are mapped onto the predictions. eHMM substantiated the presence of a long loop region present at the extracellular site of the membrane. Numerous detected SMFS signals imply the existence of a structurally complex domain in this region. These findings can contribute in understanding the mechanisms of HP_0565 that contribute to the pathogenicity triggering.

The noel approach eHMM for analyzing membrane proteins in case of HP_0565

Anne-Marie Pflugbeil, Florian Heinke, Nora Heinig and Dirk Labudde

Hochschule Mittweida, University of Applied Sciences
Technikumplatz 17, D-09648 Mittweida, Germany
email: forename.surname@hs-mittweida.de

Abstract

Membrane proteins are essential elements in metabolic pathways and play a main role in current protein research. But just a few mechanisms and structures are clarified to date. Single molecule force spectroscopy (SMFS) allows detecting molecular protein stability. In this work we present analyzed SMFS experiments and a new energy profile based description of a putative transmembrane protein of *Helicobacter pylori*, called HP_0565. This protein is assumed to be a main virulence factor of Helicpobacter pylori. To this date, there are no structural information of HP_0565 in databases.
We applied SMFS data from HP_0565 to gather information about possible existing unfolding events that are corresponding to the secondary structure elements. The prediction was performed by existing prediction methods, such as TMHMM2.0, HMMTop, MEMSAT3, ConPred II and SOSIUI. In this work we predicted a so called energy profile by the sequence of HP_0565. An energy profile is a schematic plot of the interaction energy of each residue as a function of the residue position in the sequence. Actually these residue-residue interaction energies are calculated by coarse grained models derived by known protein structures. Because of missing three-dimensional structural information, we used a novel GOR algorithm based prediction method (eGOR) which calculates an energy profile by the amino acid sequence. Furthermore, we developed a hidden Markov model (eHMM) which predicts outer- and inner membrane regions by energy profiles. Tested on an evaluation set of 110 known membrane protein structures, our membrane region prediction method showed an accuracy of 77%. The eHMM was applied to predict membrane regions by the energy profile of HP_0565.
These predicted regions showed strong correlations to predictions made by the known methods mentioned above. By a statistical analysis of unfolding events observed in force curves, we were able to classify unfolding pathways. Using information achieved by the secondary structure of HP_0565, the mapping of unfolding barriers to structure elements derived from experimental data, was realized by applying these observations to results gathered from our new eHMM approach. We used this mapping to cluster the observed unfolding pathways to main- and side pathways of HP_0565.

The eGOR algorithm and eHMM are available at http://bioservices.hs-mittweida.de/Epros.

The novel approach eHMM for analyzing membrane proteins in case of HP_0565

Anne-Marie Pflugbeil, Florian Heinke, Nora Heinig and Dirk Labudde
University of Applied Sciences, Technikumplatz 17, D-09648 Mittweida

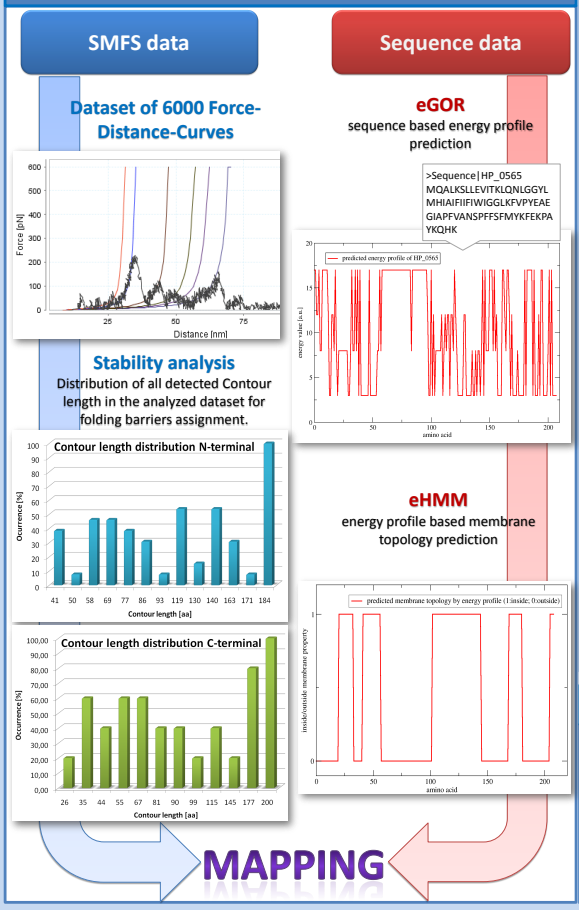
HOCHSCHULE MITTWEIDA UNIVERSITY OF APPLIED SCIENCES

Background and Motivation

Membrane proteins are essential elements in metabolic pathways and play a main role in current protein research. But just a few mechanisms and structures are clarified to date. Single Molecule Force Spectroscopy (SMFS) allows detecting molecular protein stability [1].

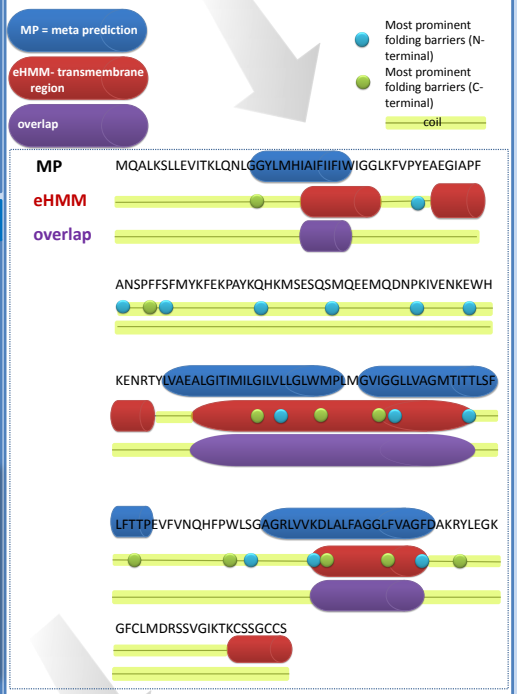
In this work we present analyzed SMFS experiments and a new energy profile based description (eGOR) of a putative transmembrane protein of *Helicobacter pylori*, called HP_0565. Furthermore we demonstrate a new approach (energy profile based Hidden Markov Model, eHMM) [3] for analyzing HP_0565. This protein is assumed to be a main virulence factor of *Helicobacter pylori*. To this date, there are no structural information of HP_0565 in data bases [2].

Methods



Results and Application

The **eHMM** was applied to **predict transmembrane regions** by the energy profile of **HP_0565**. Our membrane region prediction method showed an **accuracy of 77%**, based on PDBTM. Using information achieved by the secondary structure of HP_0565, the mapping of folding barriers to structure elements was realized enhanced by eHMM.



The **eHMM predicted** transmembrane regions (red timbers) showed strong correlations to the **meta prediction** of HP_0565 (blue timbers).

By a statistical analysis of unfolding events observed in force curves, we were able to classify unfolding pathways. The demonstrated folding barriers correlate significantly with transmembrane regions predicted by eHMM.

In conclusion the eHMM leads to an **enhanced meta prediction**.

References

[1] F. Dressel, A. Marsico, A. Tuukkanen, M. Schroeder and D. Labudde; Understanding of SMFS barriers by means of energy profiles; Proc. GCB (2007)
[2] A.M. Pflugbeil; Stabilitätsanalyse des Membranproteins HP0565- Eine experimentelle und bioinformatische Studie; Bachelor Thesis (2010)
[3] F. Heinke; Energieprofilbasierende Methoden zur Analyse von Proteinfamilien; Bachelor Thesis (2010)

eGOR algorithm and eHMM are availible at:
bioservices.hs-mittweida.de/Epros

contact: apflugbe@hs-mittweida.de

2.3.4 Analysis of membrane protein stability in nephrogenic diabetes insipidus by multiple energy profile alignment approach, MEPAL

This poster has been presented at the German Conference on Bioinformatics 2011. It summarizes the findings presented in the corresponding book chapter and paper. It mainly focuses on the theoretical aspects of the MEPAL algorithm and the findings which can be derived from its application. The analyses of membrane proteins involved in nephrogenic diabetes insipidus are discussed as a case study.

GCB 2011 – Poster Abstract

Analysis of membrane protein stability in nephrogenic diabetes insipidus by multiple energy profile alignment approach, MEPAL

Florian Heinke[1], Anne Tuukkanen[1,2], Dirk Labudde*[1]

[1] University of Applied Sciences Mittweida, Technikum Platz 17, 09648 Mittweida, Germany

[2] Technical University of Dresden, Biotechnology Center, Tatzberg 47/49, 01309 Germany

Email: Dirk Labudde* - dirk.labudde@hs-mittweida.de;

* Corresponding author

Diabetes insipidus (DI) is a rare endocrine disorder with an incidence in general population assessed on one case per 25.000-30.000 people [1]. It is a disease characterized by polyuria and compensatory polydipsia. The underlying causes of DI are diverse and can be a central defect in which no functional arginine-vasopressin is released from the pituitary. It may be caused by defects in the kidney (nephrogenic DI, NDI) as well. Four different types of NDI are known. First, acquired NDI can originate as a side-effect of drugs with the most prominent being the antibipolar drug lithium. Second and third, autosomal recessive and dominant inheritable NDI is caused by gene mutations in the AQP2 gene [2]. Finally, mutations in the AVPR2 gene, which encodes V2R, is the cause of the X-linked inheritable form of NDI. V2R is the key player in triggering the transcellular water transport by the availability of binding to the hormone arginine-vasopressin and releasing cAMP to the water resorption cascade. Known from literature, there are about 200 mutations in this receptor, which are involved in NDI. Furthermore, mutations in the aquaporin proteins, which realize the transcellular water transport in the V2R triggered cascade, lead to the loss of transport activity [3].

By our new methods, which are based on so called energy profiles, we analyzed the correlation of mutations and functionality in the V2 receptor. Additionally, we applied this approach to the aquaporins with respect to protein mechanisms. Energy profiles are derived by a novel coarse grained energy model based on statistical physics [4, 5]. As an abstraction of chemical and structural protein properties, an energy profile can be interpreted as a fold and sequence specific representation. Thus, energy profiles lead to the opportunity to compare and detect energetic divergences induced by mutations.

For a comparative analysis of influences of mutations in the proteins involved in NDI, we developed a multiple energy profile alignment algorithm (MEPAL). Based on the classic CLUSTAL approach [7], the MEPAL algorithm performs pair wise energy profile alignments and derives a distance matrix using the distance score null-hypothesis. The distance score gives a hint for alignment significance. By applying the UPGMA-algorithm, a guide/distance tree is produced and a progressive multiple energy profile alignment can be calculated. Furthermore, the consensus profile and alignment position specific energy conservations can be derived [6, 7].

In our work, we show that energetic divergences and conservations detected by MEPAL correlate with observations given by functional experiments [8]. We found evidence of reduced water flux in aquaporin-2 by aligning the energy profiles of aquaporin-2-wt and mutants given by literature [9]. Furthermore, we aligned the energy profiles of V2R in bound and unbound state with arginine-vasopressin. Detected energetically divergent regions (see

Fig.1) correspond to residues involved in hormone binding, indicating the energetic flexibility of these amino acids which is necessary for the proteins binding mechanism. These results confirm experimental observations [10]. Mutating these residues leads to the loss of hormone affinity in V2R as found in NDI.

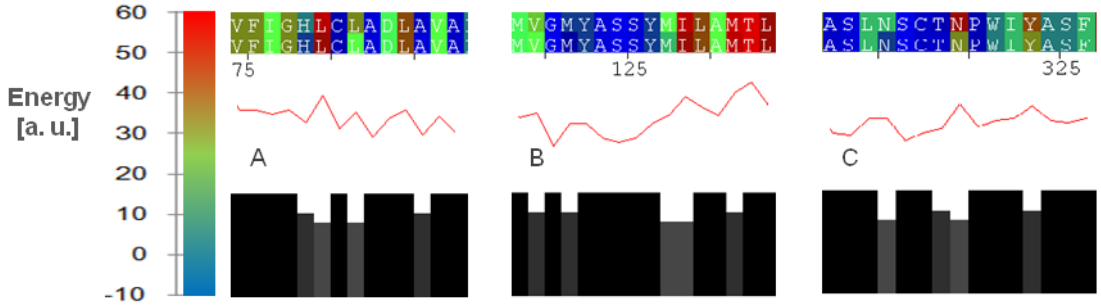


Figure 1: MEPAL output for the energy profile alignment of V2R in bound and unbound state.
Energy profiles are represented by the coloring scheme in the upper row (blue: low energy, red: high energy, green: intermediate energy). The consensus profile and energy conservation are shown in the middle and bottom row, respectively.
The energetically divergent regions (A, B, C) correspond to residues involved in hormone binding, indicating their energetic flexibility which is necessary in the V2R hormone binding mechanism.

Hence, our theoretical approach emphasized the mechanisms of the proteins involved in NDI which are described by literature and demonstrates the possibilities of this novel approach for analyzing correlations in protein evolution and functionality.

References

1 S. Ananthakrishnan. *Diabetes insipidus in pregnancy: etiology, evaluation, and management*. Endocr Pract, 15(4):377-382, 2009.

2 S. M. Mulders et al. *An aquaporin-2 water channel mutant which causes autosomal dominant nephrogenic diabetes insipidus is retained in the golgi complex*. J Clin Invest, 102(1):57 66, Jul 1998.

3 J. H. Robben, N. V. A. M. Knoers, and P. M. T. Deen. *Cell biological aspects of the vasopressin type-2 receptor and aquaporin 2 water channel in nephrogenic diabetes insipidus*. Am J Physiol Renal Physiol, 291(2):F25-F270, Aug 2006.

4 F Dressel, A Tuukkanen, M Schroeder, D Labudde. *Understanding of SMFS barriers by means of energy profiles*. Proc. GCB., 2007.

5 D. H. Wertz and H. A. Scheraga. *Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule*. Macromolecules, 11(1):9 15, 1978.

6 D. Gusfield. *Efficient methods for multiple sequence alignment with guaranteed error bounds*. Bull Math Biol, 55(1):141-154, Jan 1993.

7 D.G. Higgins, J.D. Thompson, T.J. Gibson. *Using CLUSTAL for multiple sequence alignment*. Methods Enzymol, 266, 383-402,1996-

8 N. Chakrabarti, B. Roux, and R. Pomes. *Structural determinants of proton blockage in aquaporins*. J Mol Biol, 343(2):493-510, Oct 2004.

9 B. Ilan, E.Tajkhorshid, K. Schulten, and G. A. Voth. *The mechanism of proton exclusion in aquaporin channels*. Proteins, 55(2):223-228, May 2004.

10 C. Barberis, B. Mouillac, and T. Durroux. *Structural bases of vasopressin/oxytocin receptor function*. J Endocrinol, 156(2):223-229, Feb 1998.

Analysis of membrane protein stability in *Diabetes insipidus*

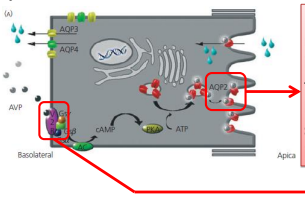
Florian Heinke[1], Anne Tuukkanen[1,2] and Dirk Labudde[1]
[1]University of Applied Sciences, Technikumplatz 17, D-09648 Mittweida
[2]Biotechnology Center TU Dresden, Tatzberg 47/49, D-01307 Dresden

HOCHSCHULE MITTWEIDA UNIVERSITY OF APPLIED SCIENCES

Background and Motivation

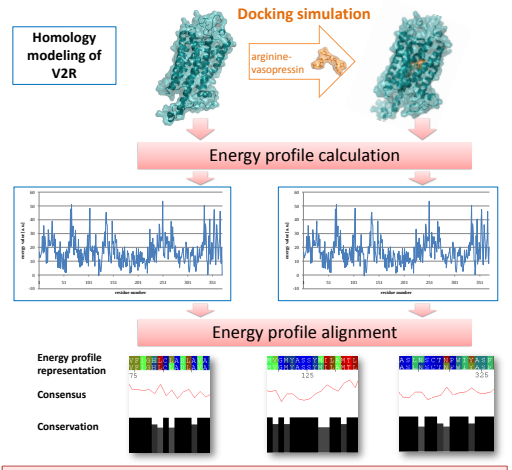
Diabetes insipidus (DI) is a rare endocrine disorder characterized by polyuria and compensatory polydipsia. It may be caused by defects in the kidney (nephrogenic DI, NDI). One known type of NDI is induced by mutations in the aquaporin-2 encoding gene and mutations in the AVPR2 gene, which encodes the arginine-vasopressin receptor V2R. These proteins are the keyplayers in transcellular water transport [1,2].

In our work we demonstrate new methods based on so called energy profiles. Our approach and developed tools (http://bioservices.hs-mittweida.de) deal with the influence of mutations on protein stability [5]. By these implemented algorithms we are able to detect and describe destabilizing influences induced by mutations in V2R and aquaporin-2. By that, it is possible to discuss the linkage between function or loss of function and mutations in proteins involved in Diabetes insipidus. We analyzed known mutations by energy profiles in the V2 receptor and the involved aquaporins with respect to protein mechanisms.



Mutations in the V2 receptor and aquaporin-2 induce the loss of water transporting functionality leading to the polyuria syndrome in Diabetes insipidus

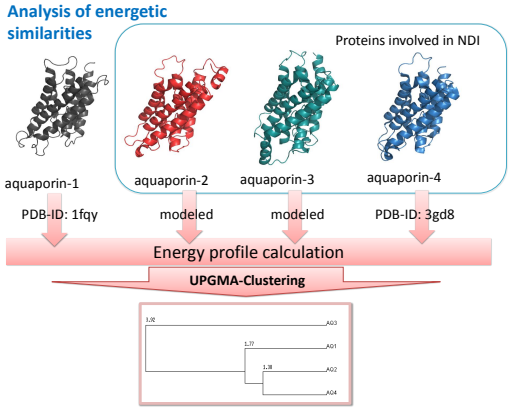
Stability analysis of V2R



The energy profile alignment shows main energy profile divergences at residues A84, I130 and P322. From literature it is known that mutating these residues leads to loss of binding ability [3].

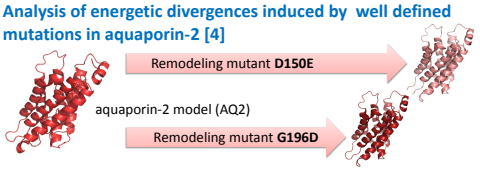
Stability analysis of aquaporins

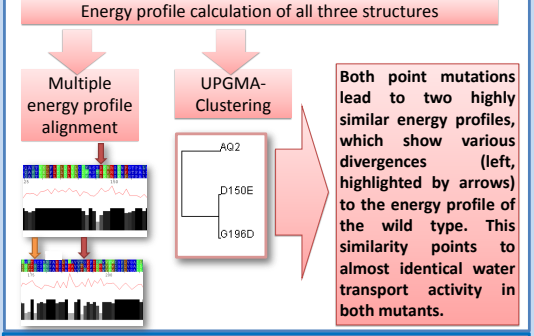
Analysis of energetic similarities



Referring to known stability information of aquaporin-1 and the energetic similarities between the involved aquaporins, **we postulate that all aquaporins show similar folding and stability characteristics leading to almost identical water flux.**

Analysis of energetic divergences induced by well defined mutations in aquaporin-2 [4]





Both point mutations lead to two highly similar energy profiles, which show various divergences (left, highlighted by arrows) to the energy profile of the wild type. This similarity points to almost identical water transport activity in both mutants.

References

[1] Deen et al. Requirement of human renal water channel aquaporin-2 for vasopressin-dependent concentration of urine. *Science* 1994; 264: 92–95.
[2] Mulders et al. An aquaporin-2 water channel mutant which causes autosomal dominant nephrogenic diabetes insipidus is retained in the Golgi complex. *J Clin Invest* 1998; 102: 57–66.
[3] UniProt entry of V2R (UniProt-ID: P30518), www.uniprot.org/uniprot/P30518
[4] Guyon et al. Characterization of D150E and G196D aquaporin-2 mutations responsible for nephrogenic diabetes insipidus: importance of a mild phenotype. *Am J Physiol Renal Physiol*. 2009;297(2):F489-98.
[5] Heinke F. Energieprofilbasierende Analysemethoden von Proteinfamilien, Bachelor thesis, Mittweida (2010)

contact: florian.heinke@hs-mittweida.de

2.3.5 eGOR - Predicting the total potential energy of a protein's native state by sequence

The prediction of the total potential energy of a structurally unknown protein of interest from its sequence can aid in the definition of folding constraints in ab initio folding simulations. In ab initio folding, a protein structure is predicted by simulating the folding process. However, this approach is computationally demanding and often leads to non-native structures, since energetic folding barriers are unknown. An improved version of eGOR can provide the identification of such folding barriers. Non-native structures occur due to limitations in simulating quantum-physical processes which prevent the gradient decent in energetic minimization. Thus, non-native structures obtained from ab initio folding show increased total potential energies as observed in the native state. The eGOR algorithm provides the prediction of an energy profile and due to correlation to physics-based potentials (see Figure 1.1) the total potential energy of the protein in its native state can be predicted. Observed discrepancies of total potential energies of simulated structures to total potential energies of structures resulting from ab initio folding point to insufficient simulated protein folding.

eGOR - Predicting the total potential energy of a protein's native state by sequence

Florian Heinke*, Steffen Grunert and Dirk Labudde

Hochschule Mittweida, University of Applied Sciences
Technikumplatz 17, D-09648 Mittweida, Germany
email: florian.heinke@hs-mittweida.de, steffen.grunert@hs-mittweida.de, dirk.labudde@hs-mittweida.de

Abstract

Motivation: The structure determination of proteins is an essential step in drug design and modern biology. However, experimental structure determination is a time- and resource-consuming procedure. *In silico*-driven techniques, such as comparative modelling or *ab initio*-folding are constricted methods as well by being limited to modelling templates or due to limited computational feasibility. A further restriction of *ab initio*-folding is that observed convergence of the protein's total potential energy (TPE) during simulation does not necessarily correspond to the protein's native state. In this work, we demonstrate a method for predicting the TPE of a given protein in its natively folded stated based on its sequence.

Materials & Methods: We derived a coarse-grained energy model which describes the interaction energy of each residue with respect to its spatial adjacent residues in a given protein structure. The sums of all coarse-grained energies of all investigated proteins correspond very well to TPEs derived by all-atom molecular dynamics (MD) analyses ($r^2 = 0.95$).
We adapted the GORIII secondary structure prediction algorithm for predicting discretized coarse-grained energies based on a protein's sequence (eGOR). By applying the correlation elucidated above to the sum of discretized coarse-grained energies predicted by eGOR, the TPE of the protein is predicted.

Results: We applied our prediction algorithm to 220 protein structures. Predicted TPEs and TPEs derived from MD correlate very well ($r^2 = 0.96$).

Conclusion: The algorithm presented in this work is suitable for predicting the TPE of a protein in its native fold based on its sequence. Thus, it can be applied as a convergence criterion in *ab initio*-folding and is helpful for distinguishing local energy minima from the global energy minimum in spatial folding trajectories of a protein of interest.

*Author to whom correspondence should be addressed

eGOR

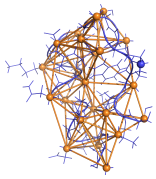
Predicting the total potential Energy of a Protein's native State by Sequence

Florian Heinke, Steffen Grunert and Dirk Labudde
University of Applied Sciences, Technikumplatz 17, D-09648 Mittweida

Background and Motivation

- **Established methods for predicting/modeling protein structures:**
  - **comparative (homology) modeling**
  - **fragment library-based modeling (i.e. MODELLER, ROBETTA)**
  - ***ab initio* folding [1,6]**
- **Drawbacks of these methods are:**
  - **necessity of structural template(s)**
  - **computational and time-demanding**
  - **energy guidance values of the native state are unknown [1,3,6]**

Methods
Computing Protein Energy Profiles



Residue-residue interactions in trpCage (PDB ID: 2JOF)

Residue-residue interaction

$$C(i,j) = \begin{cases} 1, & \|r_i - r_j\| \leq 8\text{Å} \\ 0, & \text{else} \end{cases}$$

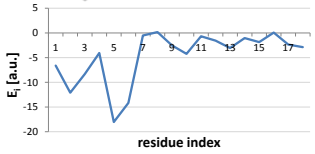
Coarse-grained energy

$$e_i \propto -\ln\left(\frac{n_{i,\text{in}}}{n_{i,\text{out}}}\right)$$

(1) $e_{i,j} = C(i,j)(e_i + e_j)$ (2) $E_i = \sum_{\langle i,j \rangle} e_{i,j}$

(3) $E_{Tot} = \sum_{i \in \text{Protein}} E_i$

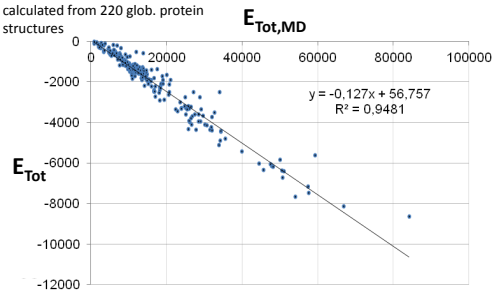
Energy profile of trpCage



[3,5]

Correlation to total potential energies derived from molecular dynamics [4]

- calculated from 220 glob. protein structures



Methods
Predicting Protein Energy Profiles

Protein Sequence

```
>aProteinSequence
NLYIQWLKDGGPSSGRPPPSILMNNHGFFTQW
```

eGOR

- adapted GOR algorithm [2]
- prediction of discretized energy profiles
- based on information theory

- **Correlation: $E_{tot,MD} \sim E_{tot,predicted}$ $R^2 = 0.96$**

Energy profile of trpCage

- predicted from sequence
- calculated from structure

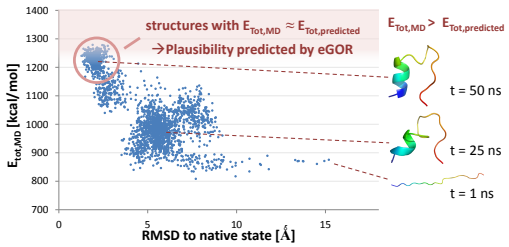
E_{Tot} can be predicted from sequence accurately

$E_{Tot,MD}$ = 1,111 kcal/mol

$E_{Tot,predicted}$ = 1,266 kcal/mol

Application and Conclusions

Case study: *Ab initio* folding simulation of TrpCage



Conclusions

- energy profiles can be predicted from sequence by eGOR
- from these, total pot. energy values E_{Tot} can be derived
- predicted E_{Tot} values correspond to E_{Tot} values observed in known protein structures
- **potential method for**
  - **predicting E_{Tot} of an unknown protein structure**
  - **protein structure assessment and concluding physical plausibility**
  - **deriving guidance values for *ab initio* folding and protein structure modeling**

References

1 Peter L. Freddolino, Feng Liu, Martin Gruebele, and Klaus Schulten. Ten-microsecond molecular dynamics simulation of a fast-folding WW domain. *Biophys J*, 94(10):L75-L77, May 2008.
2 J. Garnier, J. F. Gibrat, and B. Robson. GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol*, 266:540-553, 1996.
3 Florian Heinke and Dirk Labudde. Membrane protein stability analyses by means of protein energy profiles in case of nephrogenic diabetes insipidus. *Comput Math Methods Med*, 2012:790281, 2012.
4 J. Ponder. TINKER - software tools for molecular design. Technical report, Dept. of Biochemistry and Molecular Biophysics, Washington University, School of Medicine, St. Louis, 2001.
5 S. Tanaka and H. A. Scheraga. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, 9(6):945-950, 1976.
6 M. Zvelebil and J.O. Baum. Understanding Bioinformatics. Garland Science,first edition, 2008.

contact: florian.heinke@hs-mittweida.de dirk.labudde@hs-mittweida.de

2.4 Publications submitted for Peer-review

2.4.1 Functional Analyses of Membrane Protein Mutants involved in Nephrogenic Diabetes insipidus: An Energy-based Approach

This book chapter has been written after receiving an invitation from iConcept Press. As in the previous publications, this chapter focuses on the membrane proteins that are key players in the formation of symptoms typical for nephrogenic diabetes insipidus. However, a novel method for computing energy profiles of α-helical membrane protein structures has been introduced which incorporates topological data of the protein structure to model interaction potentials observed between residues located in the membrane bilayer more realistically. By that, the sensitivity of the approach has been increased. Furthermore, the book chapter introduces detailed SMFS data obtained from experiments on aquaporin-1. Accounting energy profile and SMFS data as well as information derived from residue-residue contacts, predictions concerning stabilizing properties in aquaporin-2 observable by means of SMFS are made. Additionally, energy profile similarities are discussed in a rather global, dScore-based way, which lead to the prediction of causes that account for the inactivation of N68S aquaporin-2 mutants. Finally, Asp 85 is proposed as a main trigger in structural rearrangements of V2R after binding to its ligand. This assumption is made based on theoretical assumptions and is substantiated by experimental findings.
The manuscript of the full book chapter presented in this section is submitted and currently in the peer-review process.

Functional Analyses of Membrane Protein Mutants involved in Nephrogenic Diabetes insipidus: An Energy-based Approach

Florian Heinke & Dirk Labudde
Department of MNI,
University of Applied Sciences Mittweida, Germany

1 Introduction

Integral membrane proteins are coded by 20-30% of all open reading frames of known genomes (Marsico et al., 2007; Brito & Andrews, 2011; Tan et al., 2008). As elements in accomplishing numerous molecular processes, i.e. signal transduction, passive and active transport of an extensive number of chemical compounds and ions, mutations in genes coding for membrane proteins are often linked to diseases (Luckey, 2008). Despite their biological importance, relatively little is known about protein folding, functional mechanics and synthesis of membrane proteins (Marsico et al., 2007). This is due to experimentally costly and complex procedures since membrane proteins are difficult to handle in lab experiments (Sadowski et al., 2008). To understand correspondences between genetic mutations and the effects on protein mechanics, the development of novel theoretical approaches is highly demanded. In our work we demonstrate a theoretical approach to discuss the influences of genetic mutations in membrane proteins which are directly linked to nephrogenic diabetes insipidus.
Nephrogenic diabetes insipidus (NDI) is a disorder which can be acquired as a side effect of surpassing drug taking or which is caused by inherited genetic mutations. Autosomal recessive and dominant inherited NDI are linked to mutations in genes encoding the integral membrane aquaporin-2 water channel (Deen et al., 1994; Mulders et al., 1998). X-linked inheritable NDI is caused by mutations in the gene encoding the AVP type-2 receptor membrane protein (V2R) (van den Ouweland et al., 1992; Rosenthal et al., 1992). In the general population, inherited NDI shows a low prevalence of one case per 20,000-30,000 people (Ananthakrishnan, 2009; Krysiak et al., 2010; Robertson, 1995). Aquaporin-2 water channels and V2R are essential elements in the water reabsorption through the apical cell membrane. This water composes the main part of pro-urine; a product that results from ultra-filtration in the kidney. The process of water reabsorption from the pro-urine is essential to ensure the body's fluid balance and is realized by membrane-integrated aquaporin-2 water channels. The insertion of aquaporin-2 into the human kidney cell membrane is triggered by the antidiuretic hormone, which is also referred to as argenine vasopressin (AVP). The AVP blood concentration is regulated by the controlled release of AVP in the pituitary gland which is adapted according to the body's fluid balance. In the process, the binding of AVP to V2R leads to the activation of the receptor. In this state V2R is able to interact with the guanine nucleotide-binding G(s) subunit alpha (Wettschureck & Offermanns, 2005; Milligan & Kostenis, 2006). Subsequently, the activation of adenyl cyclase 6 takes place and cAMP is released into the cell plasma (Defer et al., 2000; Hanoune et al., 1997). By means of protein kinase A, cAMP triggers the phosphorylation of aquaporin-2 molecules which

are stored in cytoplasmic vesicles that have bound to the endoplasmic reticulum. The phosphorylation induces the translocation and fusion of the cytoplasmic vesicles into the plasma membrane and finally leads to the insertion of aquaporin-2 molecules into the apical membrane (Kanehisa & Goto, 2000).
Inactive mutants of V2R and aquaporin-2 cause a reduced water reabsorption in the kidneys (Los et al., 2010). Consequences are the typical symptoms of NDI, e.g. sensorineural deafness, urinary tract anatomy, ataxia, peripheral neuropathy, mental retardation, psychiatric illness, a daily output of 15-20 l of highly dilute ($<$ 100 mOsmol / kg) urine (polyuria) and compensatory excessive liquid intake (Los et al., 2010; Strom et al., 1998; Birnbaumer, 2002). In newborn infants, NDI is characterized by dehydration symptoms, irritability, poor feeding as well as poor weight gain. A schematic illustration of these molecular coherences is given in Figure 1. The

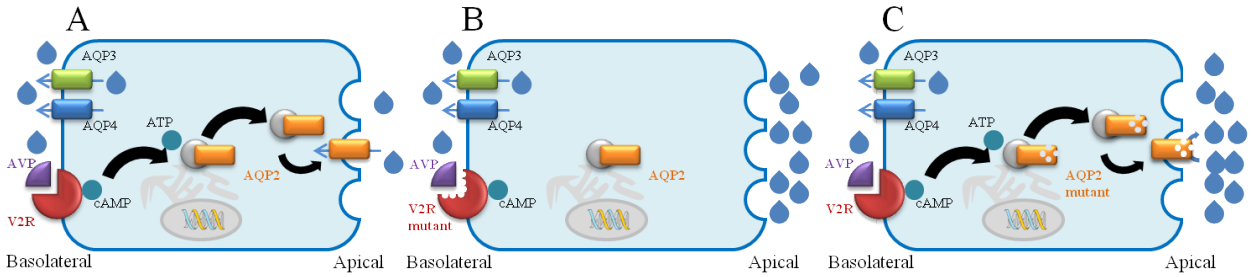


Figure 1: A: In normally regulated water absorption in kidney cells, the antidiuretic hormone arginine vasopressin (AVP) is released in the pituitary gland, binds to the V2 receptor (V2R), and subsequently induces a series of phosphorylation reactions which lead to the insertion of aquaporin-2 water channels in the apical membrane that allow water molecules to pass the membrane. B: Genetic mutations in the gene encoding V2R lead to reduced binding affinity and protein stability in V2R. Dysfunctional V2R mutants cause a significantly reduced amount of inserted aquaporin-2 proteins and thus decrease the water flux through the apical membrane. On the other hand, dysfunctional aquaporin-2 mutants decrease the water reabsorption as well (see C). Reduced water reabsorption is directly linked to an increased output of highly dilute urine (polyuria) and excessive drinking (polydipsia) which are the most severe symptoms observable in nephrogenic diabetes insipidus patients (Los et al., 2010; Robertson, 1995; Birnbaumer, 2002).

direct inspection of the aquaporin-2 gene as well as the V2 receptor gene (AVPR2) has become accomplishable in clinical practice (Fujiwara & Bichet, 2005) for differential NDI diagnosis and has been substituting dehydration testing over the last years (Los et al., 2010).
The analyses elucidated in this work focus on the stability of aquaporin-2 and stability discrepancies which can be observed in aquaporin-2 mutant structures. Further discussions deal with V2R and the shifts in stability induced by interactions with AVP. The key methodology utilized in the investigations is based on the calculation and comparison of so-called protein energy profiles. Protein energy profiles are representations of structure stability, whereas physico-chemical properties and spatial information are abstracted to a sequence of fuzzy numbers utilizing a coarse-grained energy model. This approach provides in general the opportunity to inspect spatial and conformational modifications as consequences of protein-environment interactions (Heinke & Labudde, 2012). By the pairwise energy profile comparison energetic discrepancies become observable; whereby information of effects in function and activity, which, for example, might have resulted from mutations, can be gained. More detailed discussions of energy profile-based methods are given in the following section.

2 Protein Energy Profiles as an Investigation Methodology for Protein Functionality and Stability

2.1 Theory of Protein Energy Profiles

To investigate the influences of mutations on protein function and stability, a coarse-grained knowledge-based energy model has been implemented. Like in many coarse-grained energy models, smoothing of physical information is achieved by reducing system complexity (Zhang et al., 2004). In particular, coarse-grained energies are derived in our model from statistics and concepts of statistical physics as well as by applying a straight-forward residue contact function. The basic concepts which have been applied in this model are discussed in (Sippl, 1993)) and (Tanaka & Scheraga, 1975; Tanaka & Scheraga, 1976). Basically, the energy of a residue is approximated by applying Boltzmann principles to amino acid-wise observations made from a dataset of experimental protein structures to the observed residue and the residue it is interacting with. Here, the dataset for statistics generation has consisted of 380 non-redundant α-helical transmembrane protein structures that had been obtained from the Protein Data Bank of Transmembrane Proteins (PDBTM) (Tusnady et al., 2004; Tusnady et al., 2005). Subsequently, statistics have been derived by counting the number of occurrences of each amino acid i in which the observed residues are found to be exposed at the surface ($n_{i,out}$) or found to be buried ($n_{i,in}$) in the protein structure (Dressel et al., 2007). Additionally, the topological state s of the observed residue is taken into account. If i is located inside the membrane, s is assigned as TM, with $s = \text{nTM}$ otherwise. Topological data had been obtained from the PDBTM database. C_β atoms (or C_α-atoms in cases when observing glycine) are declared as spatial residue-representative points. Using these statistics, the energy of residue i can be approximated:

$$e_i = -\ln\left(\frac{n_{i,in,s}}{n_{i,out,s}}\right) + k_i \ , \tag{1}$$

with $k_i = 0$ if $s = \text{nTM}$, or

$$k_i = -\ln\left(\frac{n_{i,\text{TM}}}{n_{i,\text{nTM}}}\right) \tag{2}$$

otherwise. To approximate the total energy E_i^*, all potentials of all interacting residues are taken into account (Dressel et al., 2007; Heinke & Labudde, 2012):

$$E_i^* = \sum_{\forall j | j \neq i} f(i,j)\,(e_i + e_j) \ , \tag{3}$$

with

$$f(i,j) = \begin{cases} 1, & \|i-j\| < 8\text{Å} \\ 0, & \text{else} \end{cases} \tag{4}$$

The sequence of all E_i^* of a given protein structure corresponds to the protein energy profile. An energy profile can be interpreted as a physiochemical and structure-specific representation, since spatial and chemical information are included in the computation. One can address that energy profile-based approaches can be realized from energy profiles derived from more sensitive, all-atom and physics-based methods (for instance see (Mrozek et al., 2006)), i.e. molecular dynamics (MD) techniques. However, the application of such techniques to membrane proteins is in general difficult to handle computationally, since the structure has to be embedded in a lipid bilayer, which increases the total number of atoms and thus system complexity. Additionally, simulations over long time scales (≈10ns – 100ns) are required to draw meaningful conclusions (Luckey, 2008). In the model applied in this study, the effects of the membrane bilayer is modelled by means of the term 2. Note that energy values computed by this model are given in arbitrary unit entities. In Figure 2, the energy profile of the modelled aquaporin-2

structure is plotted (see Figure 2C) and mapped on the protein structure (see Figure 2A and B) using a rainbow coloring scheme representing the corresponding energy values. As shown, residues with low energy values (blue) occur mainly in membrane spanning helices. In contradistinction to this observation, residues which are located in extra or intra-cellular protein regions show increased energy values, as depicted by dark yellow and green coloring. Accordingly, membrane spanning α-helices rest stabilized in the hydrophobic membrane environment. This general observation correlates to previously published data derived from experiments (for examples see (Fleming & Engelman, 2001; Finger et al., 2006; Luckey, 2008)).

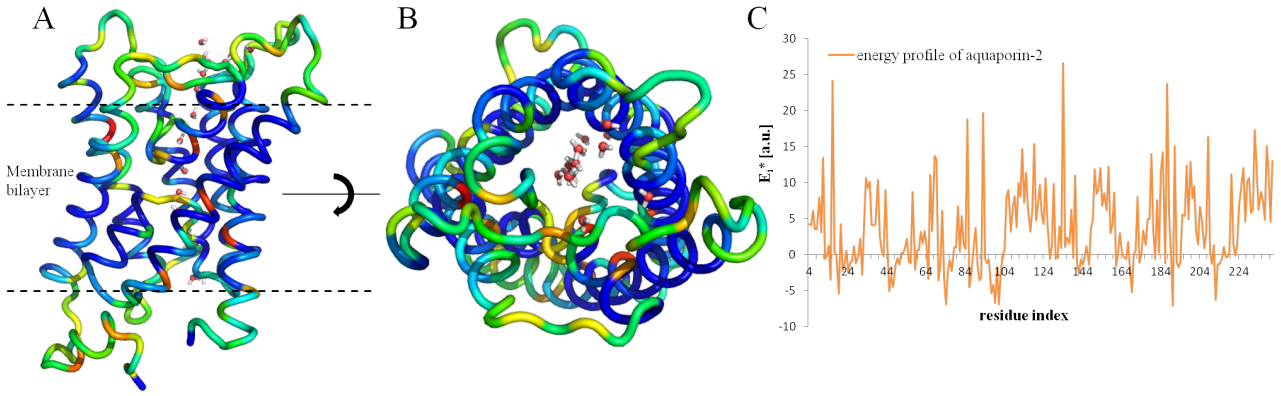


Figure 2: A and B: Energy values computed by our coarse-grained energy model are mapped onto the structure model of aquaporin-2. Obviously, residues with low energy values occur mainly in membrane spanning helices - an observation which is in agreement with experimental data (Fleming & Engelman, 2001; Finger et al., 2006; Luckey, 2008). Water molecules passing the pore are shown as spheres representation. C: The plotted energy profile of aquaporin-2.

Furthermore, a methodology for computing pairwise and multiple energy profile alignments was implemented. By that, energetic shifts can be analysed and global energy profile distances can be derived (Heinke & Labudde, 2012). Energy profile distances (dScores) can be used as input for hierarchical clustering methods, such as UPGMA (Unweighted Pair Group Method with Arithmetic Mean) (Sokal & Michener, 1958) or Neighbor-Joining (NJ) (Saitou & Nei, 1987). Similar to the approach discussed in the work of Eisenberg et al (Bowie et al., 1991) and Kozielski et al (Mrozek et al., 2006; Mrozek et al., 2007; Mrozek et al., 2009), energy profiles can be aligned by means of dynamic programming. Therefore, an energy-energy scoring function was implemented. It is derived by distances between power-equal intervals of the gaussian integral of the energy distribution. For scoring two energy values, each energy value is assigned to its interval in the gaussian integral. The distance between both integrals corresponds to the pairwise energy score. This scoring is used for aligning two given energy profiles A and B by alignment algorithms, f.e. the Needleman-Wunsch algorithm (Needleman & Wunsch, 1970) or the Smith-Waterman algorithm (Smith & Waterman, 1981). The estimation of alignment significance is provided by normalizing the resulting score x_r by taking into account the best possible score x_{opt} and the average permutation score $\bar{x}_p$. The latter is derived by permuting and realigning the given energy profiles iteratively. As discussed in (Gusfield, 1993; Higgins et al., 1996), this normalized score is referred to as distance score (dScore) and is defined as:

$$\mathrm{dScore}\,(x_r) = -\log\left(\frac{x_r - \bar{x}_p}{x_{opt} - \bar{x}_p}\right) \tag{5}$$

with

$$x_{opt}\,(A,B) = \frac{\delta(|A| + |B|)}{2}. \tag{6}$$

Here, δ denotes the best possible pairwise energy score. In general, significant energy profile alignments correspond to dScores of less than 2.5 dits. The alignment of two identical energy profiles corresponds to a dScore of 0 dits (Heinke & Labudde, 2012).

2.2 Correspondences of Energy, Function and Structure

To investigate correlations of between coarse-grained energies computed by our model and energies derived by MD, 220 globular protein structures had been obtained from the Protein Data Bank (PDB) (Berman et al., 2000) and have been analyzed by the TINKER molecular dynamics software suite ((Ponder, 2001)) subsequently. By this analysis, all-atom energies have been computed and investigated for correlations to energies derived by our coarse-grained energy model. As shown in Figure 3A, total binding energies calculated by TINKER (E_{FG}) correlate very well to the sums of all energies computed by our model (E_{CG}). Thus, our coarse-grained energy model can be used to draw physical and biological meaningful conclusions concerning residue stability and destabilizing effects of point mutations.
Furthermore, sequential, structural and functional correspondences to pairwise energy profile distances were investigated. For this purpose, 2,700 non-redundant globular protein structures and their corresponding GO-term annotations had been obtained from the PDB and UniProt (Apweiler et al., 2004), respectively. Sequence identities and structural similarities have been recorded. To investigate functional correspondences, GO-term annotations (Ashburner et al., 2000) have been compared semantically utilizing the G-SESAME web server (Du et al., 2009). As depicted in Figure 3B, sequence identities (seqId), structural similarities ($-\log(\text{p-value})$, calculated by FATCAT (Ye & Godzik, 2003)), functional similarity (semantic GO-term annotation similarity, illustrated by a blue-to-red coloring scheme) correlate strongly to energy profile distances (dScores). From this observation, it can be deduced that energy profiles yield sequential, structural and functional information as proposed. Additionally, these correspondences can be transferred to α-helical membrane proteins as well. Thus, energy profile differences correspond to functional and structural divergences and can be analysed in detail. According to this, dScores can be applied as a measure of structural stability.

3 Protein Stability of Aquaporin-2

3.1 Description of Aquaporin-2

Aquaporins belong to a family of related water channels widely present in nature. Aquaporins provide pores with high water permeability consisting of four identical subunits that form a tetramer complex during insertion in the membrane bilayer (Pollard & Earnshaw, 2007). In human tissues, 12 different isoforms are expressed but only aquaporin-2, -3 and -4 are present in the principal collecting duct cells in the membrane, whereas only aquaporin-2 is linked to NDI (King et al., 2004). Each aquaporin subunit consists of a bundle of six membrane spanning helices (helices H1-H6) and two long, distinct loop regions of each holding a short $\alpha-$helix located close to the membrane surface (helices HE, HB). As discussed in literature, two highly conserved Asn-Pro-Ala-motifs are located in both helices HE and HB, facing each other in opposite α-helical direction (Chen et al., 2006). It is proposed, that this characteristic structural feature induces a bipolar electric field which is, besides Cys189 (residue numbering according to aquaporin-1, PDB-Id 1fqy), mainly responsible for proton selectivity in aquaporins. Furthermore, it is shown that residues Phe 56, His 180 and Arg 195 induce a secondary free energy barrier located at the extracellular site of the protein which contributes to selectivity as well. An attenuation of the secondary energy barrier and, thus, reduced selectivity have been observed in Arg195 mutants by means of MD simulations (de Groot et al., 2003; Chakrabarti et al., 2004a; Chakrabarti et al., 2004b; Ilan et al., 2004). Interestingly, the residues located in the interior of the aquaporin water channels show hydrophobic properties which, as proposed,

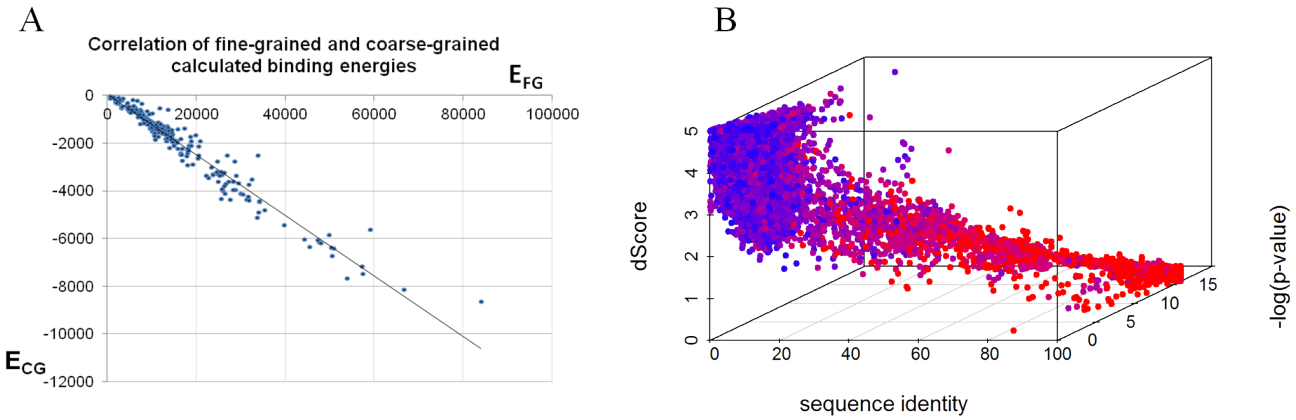


Figure 3: A: The total binding energies of 220 non-redundant globular protein structures were calculated by means of the TINKER molecular dynamics software suite (Ponder, 2001) and plotted against the sum of all E_i^* computed by our coarse-grained energy model. A linear correlation to all-atom binding energies was found. B: Scatter plot of sequence identities, structural similarities ($-\log$(p-value))and energy profile distance (dScore). Additionally, dots are colored according to the semantic similarity of the two GO-term annotations, as a representation of functional correspondences, of both proteins. Blue colored alignments indicate no detectable functional similarity between both proteins whereas red coloring points to two identical GO-term sets. As shown, dScores correspond to functional, structural and sequential similarity simultaneously. Thus, energy profile differences correspond to functional and structural divergences and can be analysed in detail.

increase water permeability. The structure of aquaporin-1 is depicted in Figure 4 with helices H1-H6 and HB with HE highlighted by orange and red coloring, respectively (see Figure 4A). In this figure, the residues mainly involved in water transport are depicted in detail in B. However, the three-dimensional structure of aquaporin-2 has not been determined experimentally yet, but, because of the strong homology in this family, it is proposed but neither experimentally nor theoretically proven that the general aforementioned aquaporin characteristics apply for aquaporin-2 as well.
Single-molecule force spectroscopy (SMFS) is one general way for investigating molecular stability and interactions experimentally. It is demonstrated in the following sections that the energy profile-based approach can be applied to transfer information derived from SMFS data to structural and stabilizing features in proteins, e.g. aquaporin-1 and aquaporin-2, which confirms the aforementioned hypothesis.

3.2 Theoretical Analysis of Protein Stability of Aquapoprin-2

Single-molecule force spectroscopy (SMFS) has been introduced as a valuable approach to investigate the stability and stabilizing effects in molecules, e.g. intra- and intermolecular forces. There have been numerous studies which discuss the use of SMFS as a method for analyzing stabilizing forces, probing energy landscapes and measuring so-called unfolding events by unfolding the membrane protein of interest in a controlled manner (Müller & Engel, 1999; Müller et al., 1999; Seelert et al., 2003; Janovjak et al., 2004; Janshoff et al., 2000). To this day, only aquaporin-1 has been investigated by SMFS (Möller et al., 2003). However, SMFS is still an error-prone, computationally- and resource-demanding methodology. Additionally, understanding and accommodating experimental observations to structural data by means of a generalized model has not been achieved yet.
As an approach for investigating protein stability in aquaporins, the energy profiles of aquaporin-1,-2,-3,-4, and -5

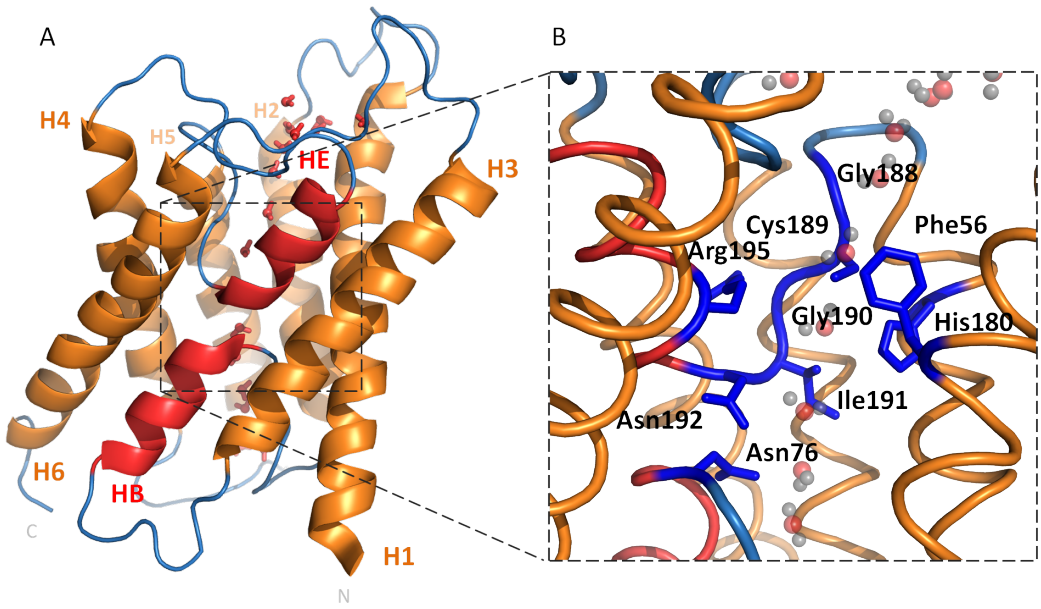


Figure 4: Homology studies of the aquaporin family suggest that all aquaporins share common structural features. It is shown, that aquaporins consist of six membrane spanning helices (H1-H6, highlighted in gold in A) and two short membrane loops, each embedding a single short helix (HB and HE, highlighted in red in A). In each helix HB and HE, a highly conserved Asn-Pro-Ala motif is present. In the folded structure, both motifs face each other in opposite direction and establish a biploar electric field which is proposed to be responsible for water permeability (de Groot et al., 2003; Chakrabarti et al., 2004a; Chakrabarti et al., 2004b; Ilan et al., 2004). B: The residues involved in water transport are highlighted in the structure by sticks representation (PDB-Id: 1fqy, residue numbering is given according to aquaporin-1).

have been computed and aligned for deriving dScores. Additionally, SMFS data of aquaporin-1 has been gathered from literature and studied for correlations with energy profile and structure data of aquaporin-1 and aquaporin-2. However, as elucidated earlier, energy profile-based analyses require spatial information of the protein of interest. To this end, the three-dimensional structures of aquaporin-1 (PDB-Id: 1fqy), aquaporin-4 (PDB-Id: 3gd8), aquaporin-5 (PDB-Id: 3d9s) have been obtained from the PDB. Theoretical structure models of aquaporin-2 and -3 have been retrieved from the ModBase database (Pieper et al., 2011). The aquaporin-2 model has been produced by utilizing comparative modelling and the protein structure of aquaporin-5 serving as modeling template (PDB-ID 3d9s). With both proteins sharing a sequence identity of 68%, the resulting structure model has been found to be reliable in quality. Quality re-evaluation has been performed by means of the protein structure analysis tool VADAR (version 1.8) (Willard et al., 2003). One measure of quality is the quality index which summarizes side chain misfoldings, stereo-chemical clashes and insufficient atom packing for each residue. Residues reported with a quality indices <4 are considered to be poorly modelled. The quality index plot of aquaporin-2 is shown in Figure 5. As illustrated, most residues in the aquaporin-2 model are found to be modelled reliably and only few spots of rather poor modelling quality are reported. In addition, the structure model of aquaporin-3 has been obtained from ModBase. Analogue to the aquaporin-2 model, this structure model has been generated by means of comparative modelling. However, only the aquaporin-homolog *E. coli* glycerol facilitator (PDB-Id: 1ldf) has been found to be the best matching experimental structure which is acceptable for comparative modelling (43% sequence identity). Subsequent comparative modelling has resulted to an aquaporin-3 model with average relia-

bility (data not shown).
In the process of analyses, the corresponding energy profiles have computed and aligned in a pairwise manner as elucidated earlier. From this, all pairwise dScores have been derived and used for hierarchical clustering by means of UPGMA. By that, a dScore-distance tree has been generated (see Figure 6). The longest branch in this tree holds a dScore of about 1.5 bits, which corresponds to twice the distance between the energy profile of aquaporin-3 and the other investigated aquaporins. However, taking into account the correspondences shown in Figure 3, this retrieved dScore points to a good agreement and similarity in energy profile progression. Hence, it can be hypothesized that stabilizing features and characteristics are conserved in the aquaporin family.

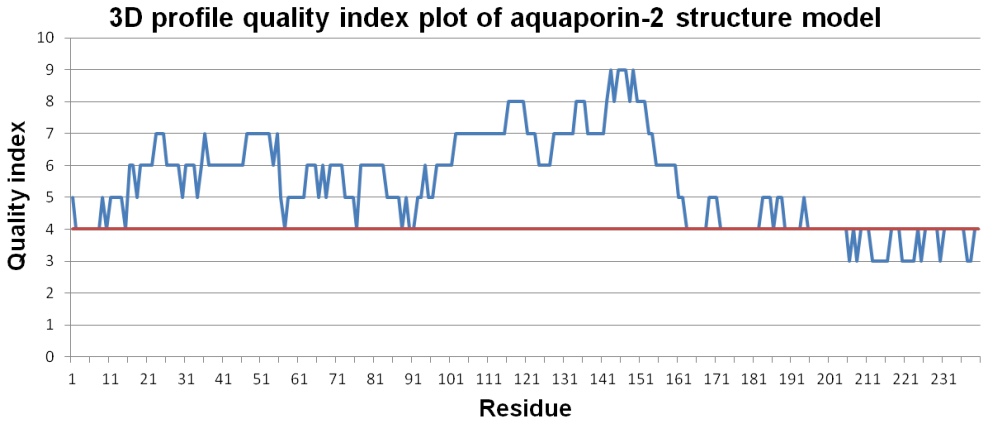


Figure 5: In energy profile-based analyses structure data of the protein of interest is required. In this study, a structure model of aquaporin-2 has been retrieved from the ModBase database (Pieper et al., 2011). Model quality has been assessed by utilizing VADAR (Willard et al., 2003). One measure of quality given by VADAR is the residue quality index which reports side chain misfolding, stereochemical overlaps and insufficient atom packing. Residues showing a quality index <4 (red line) are assessed as poorly modelled. In this case, the obtained aquaporin-2 model is found to be of good modelling quality.

To strengthen this hypothesis, we have investigated SMFS data derived from different α-helical membrane proteins concerning possible correlations to energy profiles. We found, that unfolding events can be identified on the level of energy profiles by taking into account the residue-residue contact information that can be computed from the given protein structure. As discussed in the previous section, a contact between to residues i and j is assumed according to Equation 4). The information needed for further investigations is generated by computing the sum of contacts c_i for each residue i. In the following argumentations the sequence of residue-residue contact sums $(c_1, c_2, \ldots, c_i, \ldots, c_n)$ is referred to as the residue-residue contact profile (RRCP). In general, an RRCP is a representation of spatial information. The unfolding events and stabilizing effects, that can be detected by utilizing SMFS, are manifested in spatial and physo-chemical features. Thus, taking into account the spatial information yield by RRCPs as well as the information yield in energy profiles, these features can be identified.
Regarding to aquaporin-2, it can be observed that the RRCPs of aquaporin-1 and aquaporin-2 share common features and are highly similar. With respect to known SMFS and energy profile data, it can be postulated that SMFS data derived from aquaporin-1 can be directly transferred to aquaporin-2, which emphasizes the affinity of protein stability in both aquaporins. Experiments carried out on aquaporin-2 might very likely result to similar observations. According to this and the results given in the previous paragraph, stabilizing features are conserved in the aquaporin family. To depict these agreements, the RRCPs and energy profiles of aquaporin-2 and aquaporin-1 as well as SMFS data of aquaporin-1 (e.g. observed SMFS peaks) are illustrated in Figure 7A. By transferring

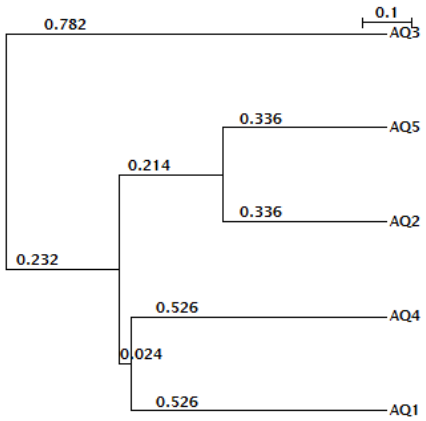


Figure 6: The three-dimensional structures of aquaporin-1, -4 and -5 had been retrieved from the Protein Data Bank and structure models of aquaporin-2 and -3 had been generated. From the pairwise energy profile comparisons of these structures, dScores have been computed and used for generating a dScore-distance tree by means of Unweighted Pair Group Method with Arithmetic Mean (UPGMA) as depicted in this figure. Short branch lengths indicate high energy profile similarities. Since energy profile similarity correlates to common stabilizing and functional features in protein structures, it can be proposed that this observation can be applied to the investigated aquaporins as well. We postulate that energetic and, thus, stabilizing characteristics are highly conserved in aquaporins. This allows the energy profile-based investigation of destabilizing effects in aquaporin-2 mutants which lead to NDI.

RRCP, energy profile and SMFS data of aquaporin-1 to aquaporin-2, SMFS unfolding events in aquaporin-2 can be predicted (represented by green bars in Figure 7B).

3.3 Protein Stability of Aquaporin-2 Mutants linked to NDI

Over the last decades, numerous aquaporin-2 mutants have been identified in NDI patients. In our analysis, we focused on seven different mutations which differ in activity and phenotype. First, the mutations D150E and L22V+C181W lead to a reduced water flux through the apical membrane (Guyon et al., 2009; Canfield et al., 1997). In contrast, the mutants T125M+G175R and G196D are experimentally affirmed to prevent water permeability totally (Goji et al., 1998; Guyon et al., 2009). Furthermore, it can be observed that mutations in the gene encoding aquaporin-2 can lead to misfolding of the protein and prevent routing of the protein to the plasma membrane. The mutants A147T and T126M have been chosen to represent this group of mutants (Mulders et al., 1997). Finally, N68S was included in the energy profile-based analysis. This mutation is located in the first conserved Asn-Pro-Ala motif. Although it is proven that the N68S mutant does not show water permeability, it is not clear whether the reduced water flux is caused by a disrupted water pore or by protein misfolding and, thus, impaired transport out of the endoplasmic reticulum (Mulders et al., 1997).
To investigate potential correspondences, the mutant structures had been generated using comparative modelling with the structure of aquaporin-5 as modelling template. Subsequently, the energy profiles had been computed, aligned and, from this, dScores have been derived. The dScore-distance tree obtained by means of UPGMA (see Figure 8) shows distinct clustering. It can be seen, that mutants leading to an impaired transport are clustered in a single group integrating the N68S mutant as well. According to this, it can be predicted that the bipolar electric field established by α-helices HB and HE is not present in N68S mutants. This affects aquaporin-2 folding significantly in such a way that the native conformation cannot be attained. This leaves N68S mutants

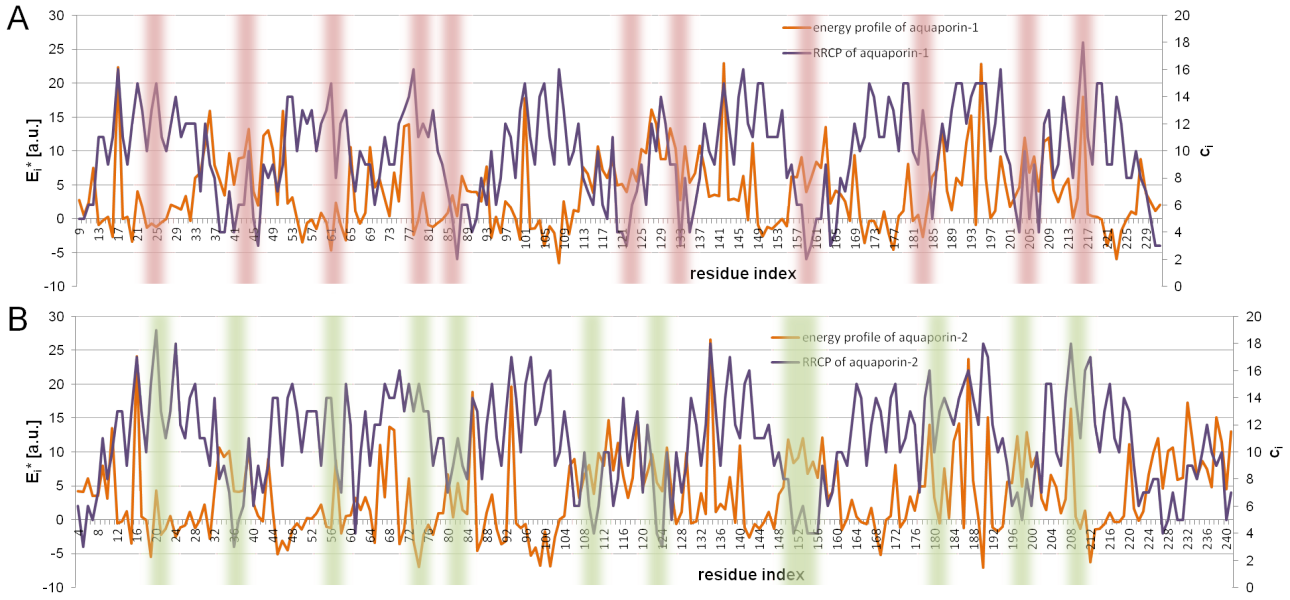


Figure 7: As given by energy profile distances and shown in Figure 6, energy profile characteristics are highly similar in the investigated aquaporin structures and, thus, conserved in the aquaporin family. In this Figure, the energy profile and residue-residue contact profile (RRCP) of aquaporin-1 (see A) and aquaporin-2 (see B) are illustrated. A residue-residue contact is assumed, if $f(i,j) = 1$ (see Equation 4). The sum of all contacts of residue i corresponds to the residue-residue contact number c_i. An RRCP corresponds to the sequence $(c_1, \ldots, c_i, \ldots, c_n)$. We investigated diverse α-helical membrane protein structures with data derived from single-molecule force spectroscopy (SMFS); an approach for measuring and probing protein energy landscapes and stabilizing characteristics (Müller & Engel, 1999; Müller et al., 1999). Correlations have been found in unfolding events (peaks) derived by SMFS experiments to energy profile and RRCP features. The positions of detected SMFS peaks in aquaporin-1 are highlighted by red bars in A. The energy profiles and RRCPs of both aquaporins share common features. Hence, the positions of SMFS peaks in aquaporin-2 are very likely similar located and distinct as observed in aquaporin-1. Predicted SMFS peaks in aquaporin-2 are highlighted by green bars in B

in a misfolded and not-active state resting in the endoplasmic reticulum. Furthermore, the mutants D150E and G196D form a separate cluster distant from the correctly folded but (partially) dysfunctional L22V+C181W and T125M+G175R mutants. This indicates the independence of occurring effects induced by these mutations. The high similarity of the energy profiles of D150E mutant and G196D mutant points to close correspondences in the mutated water transport mechanism. However, the observed energetic divergences lead to the significantly different, experimentally confirmed phenotypes. According to our hypothesis, this conclusions are analogue to the mutants L22V+C181W and T125M+G175R, where, as indicated by a branch length of 0.02, the energy profile distance and, thus, energetic divergences are much greater than in the cluster holding the D150E and G196D mutants.

Thus, the observed energy profile distances are in good agreement with experimental data and correspond to water transport activity in aquaporin-2 mutants as well as overall protein stability since mutants leading to protein misfolding can be distinguished on the level of energy profiles.

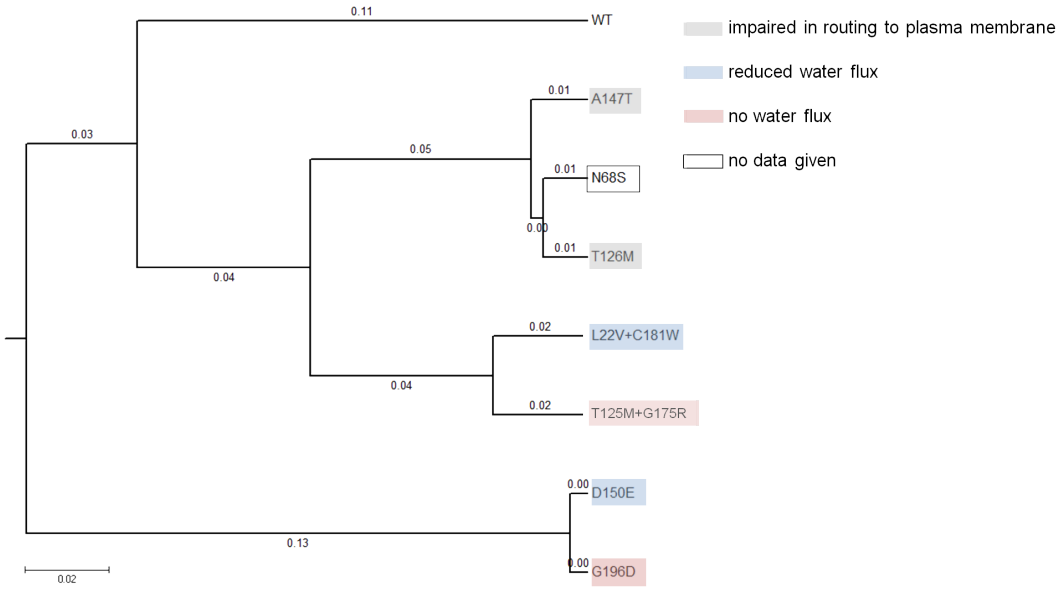


Figure 8: Structure models of seven aquaporin-2 mutants and, subsequently, corresponding energy profiles were generated. Pairwise dScores were computed and used for hierarchical clustering by means of UPGMA. As shown, groupings obtained by clustering correspond to phenotypes and functionality observed in experiments. The energy profile of the N68S mutant is arranged to the energy profiles of mutants A147T and T126M. Both mutants are proven to be responsible for incorrect folding, leaving the resulting dysfunctional protein resting in the endoplasmic reticulum. By that, no membrane insertion takes place and water reabsorption is not established leading to symptoms of NDI. From the clustering of both mutants with N68S, it can be predicted that protein structure misfolding is the most likely cause for the loss of functionality in N68S mutants. The clades of (partially) dysfunctional mutants D150E, G196D and L22V+C181W, T125M+G175R, respectively, show relatively long energy profile distances. This indicates that in each clade destabilizing effects are highly similar with differences leading to the differing phenotypes observed in experiments.

4 Investigation of Protein Stability in V2 Vasopressin Receptor

4.1 Description of V2 Vasopressin Receptor

The V2 vasopressin receptor (V2R) belongs to the class A of the so-called G-protein-coupled receptors. Like most members in this class, V2R consists of seven transmembrane α-helices (Barberis et al., 1998). The residues involved in binding the agonist anti-diuretic hormone arginine-vasopressine (AVP) are basically located in α-helices H2–H5, spanning the four sequence regions 88–96, 119–127, 284–291 and 311–317 (Slusarz et al., 2006). Once AVP has bound to V2R, the structure of V2R is passing through multiple allosteric rearrangements. In the subsequent active state, V2R is capable of interacting with cytosolic G-protein activating adenyl-cyclase. As a result of this interaction, a cascade of multiple phosporylation events is taking place, whereas, at last, aquaporin-2 is activated, translocated and finally integrated in the apical membrane (Los et al., 2010; Robben et al., 2006).
For the energy profile-based investigation of bound and unbound states of V2R, a structure model had to be computed since, to this day, no experimentally determined structure is listed in the public structure databases and, because of the limitations of comparative modelling, no reliable model had been generated also. Thus, extensive structure modelling needed to be carried out. By means of the I-TASSER modelling pipeline (Roy et al., 2010) and NAMD2 (Phillips et al., 2005) as the subsequent program of energy minimization and model-reliability assessment, a structure model of V2R has been generated. In further MD simulations, the root mean square fluctuation of the C_α-trace of the V2R structure model has been found to be 2.7 Å which confirmed the good modelling quality that had been additionally evaluated by VADAR (Willard et al., 2003). Furthermore, the docking of AVP to V2R was simulated utilizing the Molecular Docking Server (Bikadi & Hazai, 2009). Additionally, a structure model of V2R bound to AVP has been generated.

4.2 Analysis of Energy Profiles of V2 Vasopressin Receptor in bound and unbound State

From the two resulting structure models, both energy profiles have been generated and compared directly. The most significant energetic shifts ΔE_i^* (defined as $\Delta E_i^* = E_{i,\text{unbound}}^* - E_{i,\text{bound}}^*$) correspond very well to the sequence regions responsible for AVP binding (see Figure 9). Additionally, a large number of mutants have been identified which are located in these sequences regions and which negatively affect AVP binding. Well-described examples are Ala84Asp (Albertazzi et al., 2000), Ile130Phe (Pasel et al., 2000; Robben et al., 2005) and Pro322Ser (Morin et al., 1998; Vargas-Poussou et al., 1997). Especially Asp 85 shows the greatest energetic shift, e.g. an energetic increase of 8.8 a.u. during binding. As a hydrophilic and polar residue, aspartic acid is more often observed in extra- and intracellular regions of membrane proteins than in the rather hydrophobic environment of membrane spanning helices or membran-associated regions in general. With respect to the model applied in this study, the value k_{Asp} described by equation 2 is found to be >1 for aspartic acids which are located in membrane-spanning segments. Thus, interactions with membrane-located aspartic acids are approximated to cause destabilizing influences on α-helical membrane proteins in general. With focus on the bound and unbound states of V2R, the energetic increase of Asp 85 is caused by interactions to AVP. According to our hypothesis, Asp85 and the destabilization during and after coupling to AVP are main triggers for the structural rearrangements of the receptor. Previously published results from experimental studies substantiate these conclusions (Sadeghi et al., 1997). These studies indicate the responsibility of Asp 85 in coupling to guanine nucleotide-binding G(s) subunit alpha. As reported, Asp85Asn mutants show a 20-fold decrease in coupling efficiency. Thus, the application of the energy profile-based approach as a method for investigating protein structure stability is substantiated by these experimental findings.
However, further energy profile-based inspections are limited due to restrictions in modelling large-scale molecular rearrangements.

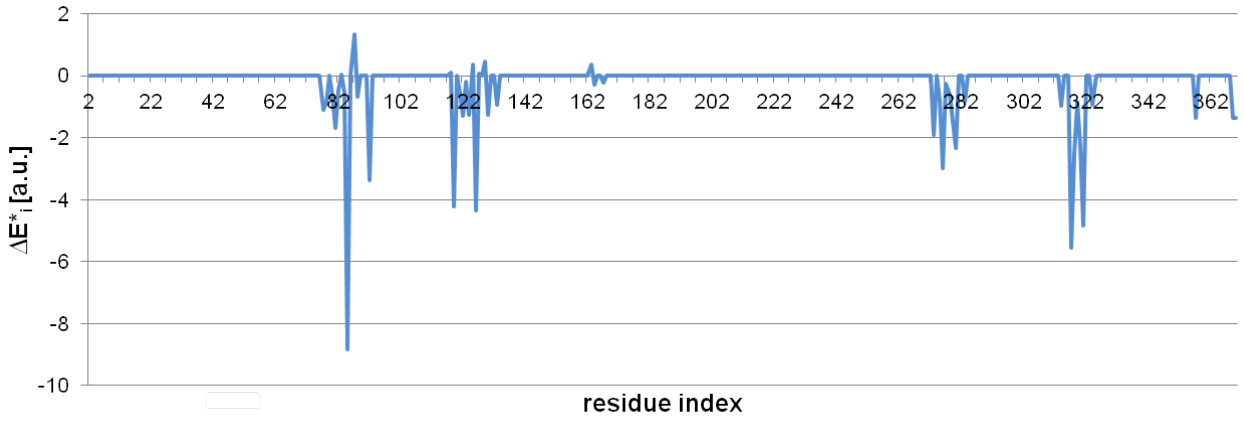


Figure 9: The generation of a human V2R structure model had been carried out by means of molecular modelling and MD. Furthermore, the docking of arginine-vasopressin (AVP) to the structure model has been performed. The superposition of the energy profiles (with $\Delta E_i^* = E_{i,\text{unbound}}^* - E_{i,\text{bound}}^*$) of V2R in bound and unbound state to AVP shows four distinct regions with varying energy values. These regions correspond to the residues mainly involved in AVP binding. The most distinct energetic shift is found at residue Asp85. From this observation, it can be postulated that the destabilization of the polar and hydrophilic Asp 85 plays a main role in the structural rearrangements observed in V2R after binding to AVP. Experimental data published by Sadeghi et al substantiate this conclusion (Sadeghi et al., 1997).

5 Conclusion

In this study membrane proteins and mutants which are involved in nephrogenic diabates insipidus have been investigated on the basis of theoretical assumptions. By this, a coarse-grained energy model has been applied which allows the calculation of so-called energy profiles from protein structure data. As shown, energy profiles can be compared and aligned to investigate discrepancies in protein structure, function and stability. In the cases of the analysed proteins and mutants, experimental observations have been substantiated by employing these approaches. Furthermore, predictions concerning effects on protein stability in protein mutants could be made by including experimental data.
Similar to the procedures elucidated in this work, analogue biological questions might be addressed in general in the same manner. Thus, efficient high-throughput in silico techniques might be established that permit the comparison to experimental data and draw valuable conclusions concerning the stability and possible stability variations of proteins of interest.

6 Acknowledgements

The authors would like to thank the Free State of Saxony and the University of Applied Sciences Mittweida for funding. Additionally, very special thanks go to Daniel Stockmann, Steffen Grunert and Michael Spranger for their support, motivation and programming.

References

Albertazzi, E. et al. (2000). Nephrogenic diabetes insipidus: functional analysis of new AVPR2 mutations identified in italian families. *J Am Soc Nephrol*, 11(6), 1033–1043.

Ananthakrishnan, S. (2009). Diabetes insipidus in pregnancy: etiology, evaluation, and management. *Endocr Pract*, 15(4), 377–382.

Apweiler, R. et al. (2004). Uniprot: the universal protein knowledgebase. *Nucleic Acids Res*, 32(Database issue), D115–D119.

Ashburner, M. et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1), 25–29.

Barberis, C., Mouillac, B., & Durroux, T. (1998). Structural bases of vasopressin/oxytocin receptor function. *J Endocrinol*, 156(2), 223–229.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Res*, 28(1), 235–242.

Bikadi, Z. & Hazai, E. (2009). Application of the PM6 semi-empirical method to modeling proteins enhances docking accuracy of AutoDock. *J Cheminform*, 1, 15.

Birnbaumer, M. (2002). V2R structure and diabetes insipidus. *Receptors Channels*, 8(1), 51–56.

Bowie, J. U., Lüthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253(5016), 164–170.

Brito, G. C. & Andrews, D. W. (2011). Removing bias against membrane proteins in interaction networks. *BMC Syst Biol*, 5, 169.

Canfield, M. C., Tamarappoo, B. K., Moses, A. M., Verkman, A. S., & Holtzman, E. J. (1997). Identification and characterization of aquaporin-2 water channel mutations causing nephrogenic diabetes insipidus with partial vasopressin response. *Hum Mol Genet*, 6(11), 1865–1871.

Chakrabarti, N., Roux, B., & Poms, R. (2004a). Structural determinants of proton blockage in aquaporins. *J Mol Biol*, 343(2), 493–510.

Chakrabarti, N., Tajkhorshid, E., Roux, B., & Pomès, R. (2004b). Molecular basis of proton blockage in aquaporins. *Structure*, 12(1), 65–74.

Chen, H., Wu, Y., & Voth, G. A. (2006). Origins of proton transport behavior from selectivity domain mutations of the aquaporin-1 channel. *Biophys J*, 90(10), L73–L75.

de Groot, B. L., Frigato, T., Helms, V., & Grubmller, H. (2003). The mechanism of proton exclusion in the aquaporin-1 water channel. *J Mol Biol*, 333(2), 279–293.

Deen, P. M., Verdijk, M. A., Knoers, N. V., Wieringa, B., Monnens, L. A., van Os, C. H., & van Oost, B. A. (1994). Requirement of human renal water channel aquaporin-2 for vasopressin-dependent concentration of urine. *Science*, 264(5155), 92–95.

Defer, N., Best-Belpomme, M., & Hanoune, J. (2000). Tissue specificity and physiological relevance of various isoforms of adenylyl cyclase. *Am J Physiol Renal Physiol*, 279(3), F400–F416.

Dressel, F., Marsico, A., Tuukkanen, A., Schroeder, M., & Labudde, D. (2007). Understanding of SMFS barriers by means of energy profiles. In *Proceedings of German Conference on Bioinformatics* (pp. 90–99).

Du, Z., Li, L., Chen, C. F., Yu, P. S., & Wang, J. Z. (2009). G-SESAME: web tools for GO-term-based gene similarity analysis and knowledge discovery. *Nucleic Acids Res.*, 37, W345–349.

Finger, C., Volkmer, T., Prodhl, A., Otzen, D. E., Engelman, D. M., & Schneider, D. (2006). The stability of transmembrane helix interactions measured in a biological membrane. *J Mol Biol*, 358(5), 1221–1228.

Fleming, K. G. & Engelman, D. M. (2001). Specificity in transmembrane helix-helix interactions can define a hierarchy of stability for sequence variants. *Proc Natl Acad Sci U S A*, 98(25), 14340–14344.

Fujiwara, T. M. & Bichet, D. G. (2005). Molecular biology of hereditary diabetes insipidus. *J Am Soc Nephrol*, 16(10), 2836–2846.

Goji, K., Kuwahara, M., Gu, Y., Matsuo, M., Marumo, F., & Sasaki, S. (1998). Novel mutations in aquaporin-2 gene in female siblings with nephrogenic diabetes insipidus: evidence of disrupted water channel function. *J Clin Endocrinol Metab*, 83(9), 3205–3209.

Gusfield, D. (1993). Efficient methods for multiple sequence alignment with guaranteed error bounds. *Bull Math Biol*, 55(1), 141–154.

Guyon, C., Lussier, Y., Bissonnette, P., Leduc-Nadeau, A., Lonergan, M., Arthus, M.-F., Perez, R. B., Tiulpakov, A., Lapointe, J.-Y., & Bichet, D. G. (2009). Characterization of D150E and G196D aquaporin-2 mutations responsible for nephrogenic diabetes insipidus: importance of a mild phenotype. *Am J Physiol Renal Physiol*, 297(2), F489–F498.

Hanoune, J., Pouille, Y., Tzavara, E., Shen, T., Lipskaya, L., Miyamoto, N., Suzuki, Y., & Defer, N. (1997). Adenylyl cyclases: structure, regulation and function in an enzyme superfamily. *Mol Cell Endocrinol*, 128(1-2), 179–194.

Heinke, F. & Labudde, D. (2012). Membrane protein stability analyses by means of protein energy profiles in case of nephrogenic diabetes insipidus. *Comput Math Methods Med*, 2012, 790281.

Higgins, D. G., Thompson, J. D., & Gibson, T. J. (1996). Using CLUSTAL for multiple sequence alignments. *Methods Enzymol*, 266, 383–402.

Ilan, B., Tajkhorshid, E., Schulten, K., & Voth, G. A. (2004). The mechanism of proton exclusion in aquaporin channels. *Proteins*, 55(2), 223–228.

Janovjak, H., Struckmeier, J., Hubain, M., Kedrov, A., Kessler, M., & Müller, D. J. (2004). Probing the energy landscape of the membrane protein bacteriorhodopsin. *Structure*, 12(5), 871–879.

Janshoff, Neitzert, Oberdörfer, & Fuchs (2000). Force spectroscopy of molecular systems-single molecule spectroscopy of polymers and biomolecules. *Angew Chem Int Ed Engl*, 39(18), 3212–3237.

Kanehisa, M. & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1), 27–30.

King, L. S., Kozono, D., & Agre, P. (2004). From structure to disease: the evolving tale of aquaporin biology. *Nat Rev Mol Cell Biol*, 5(9), 687–698.

Krysiak, R., Kobielusz-Gembala, I., & Okopien, B. (2010). Recurrent pregnancy-induced diabetes insipidus in a woman with hemochromatosis. *Endocr J*, 57(12), 1023–1028.

Los, E. L., Deen, P. M. T., & Robben, J. H. (2010). Potential of nonpeptide (ant)agonists to rescue vasopressin V2 receptor mutants for the treatment of X-linked nephrogenic diabetes insipidus. *J Neuroendocrinol*, 22(5), 393–399.

Luckey, M. (2008). *Membrane Structural Biology - With Biochemical and Biophysical Foundation*. Cambridge University Press.

Marsico, A., Labudde, D., Sapra, T., Muller, D. J., & Schroeder, M. (2007). A novel pattern recognition algorithm to classify membrane protein unfolding pathways with high-throughput single-molecule force spectroscopy. *Bioinformatics*, 23(2), e231–e236.

Milligan, G. & Kostenis, E. (2006). Heterotrimeric G-proteins: a short history. *Br J Pharmacol*, 147 Suppl 1, S46–S55.

Möller, C., Fotiadis, D., Suda, K., Engel, A., Kessler, M., & Müller, D. J. (2003). Determining molecular forces that stabilize human aquaporin-1. *J Struct Biol*, 142(3), 369–378.

Morin, D., Ala, Y., Sabatier, N., Cotte, N., Hendy, G., Vargas, R., Dechaux, M., Antignac, C., Hibert, M., Bichet, D., & Barberis, C. (1998). Functional study of two V2 vasopressin mutant receptors related to NDI. P322S and P322H. *Adv Exp Med Biol*, 449, 391–393.

Mrozek, D., Malysiak, B., & Kozielski, S. (2006). EAST: Energy Alignment Search Tool. In L. Wang, L. Jiao, G. Shi, X. Li, & J. Liu (Eds.), *Fuzzy Systems and Knowledge Discovery*, volume 4223 of *Lecture Notes in Computer Science* (pp. 696–705).: Springer Berlin / Heidelberg.

Mrozek, D., Malysiak, B., & Kozielski, S. (2007). An optimal alignment of proteins energy characteristics with crisp and fuzzy similarity awards. In *FUZZ-IEEE'07* (pp. 1–6).

Mrozek, D., Malysiak-Mrozek, B., & Kozielski, S. (2009). Alignment of protein structure energy patterns represented as sequences of fuzzy numbers. In *Fuzzy Information Processing Society, 2009. NAFIPS 2009. Annual Meeting of the North American*.

Mulders, S. M. et al. (1998). An aquaporin-2 water channel mutant which causes autosomal dominant nephrogenic diabetes insipidus is retained in the Golgi complex. *J Clin Invest*, 102(1), 57–66.

Mulders, S. M., Knoers, N. V., Van Lieburg, A. F., Monnens, L. A., Leumann, E., Wühl, E., Schober, E., Rijss, J. P., Van Os, C. H., & Deen, P. M. (1997). New mutations in the AQP2 gene in nephrogenic diabetes insipidus resulting in functional but misrouted water channels. *J Am Soc Nephrol*, 8(2), 242–248.

Müller, D. J. & Engel, A. (1999). Voltage and pH-induced channel closure of porin OmpF visualized by atomic force microscopy. *J Mol Biol*, 285(4), 1347–1351.

Müller, D. J., Sass, H. J., Müller, S. A., Büldt, G., & Engel, A. (1999). Surface structures of native bacteriorhodopsin depend on the molecular packing arrangement in the membrane. *J Mol Biol*, 285(5), 1903–1909.

Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3), 443–453.

Pasel, K., Schulz, A., Timmermann, K., Linnemann, K., Hoeltzenbein, M., Jskelinen, J., Grters, A., Filler, G., & Schneberg, T. (2000). Functional characterization of the molecular defects causing nephrogenic diabetes insipidus in eight families. *J Clin Endocrinol Metab*, 85(4), 1703–1710.

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., & Schulten, K. (2005). Scalable molecular dynamics with NAMD. *J Comput Chem*, 26(16), 1781–1802.

Pieper, U. et al. (2011). ModBase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res*, 39(Database issue), D465–D474.

Pollard, D. & Earnshaw, W. (2007). *Cell Biology*. Springer Verlag Berlin Heidelberg.

Ponder, J. (2001). *TINKER – Software Tools for Molecular Design*. Technical report, Dept. of Biochemistry and Molecular Biophysics, Washington University, School of Medicine, St. Louis.

Robben, J. H., Knoers, N. V. A. M., & Deen, P. M. T. (2005). Characterization of vasopressin v2 receptor mutants in nephrogenic diabetes insipidus in a polarized cell model. *Am J Physiol Renal Physiol*, 289(2), F265–F272.

Robben, J. H., Knoers, N. V. A. M., & Deen, P. M. T. (2006). Cell biological aspects of the vasopressin type-2 receptor and aquaporin 2 water channel in nephrogenic diabetes insipidus. *Am J Physiol Renal Physiol*, 291(2), F257–F270.

Robertson, G. L. (1995). Diabetes insipidus. *Endocrinol Metab Clin North Am*, 24(3), 549–572.

Rosenthal, W., Seibold, A., Antaramian, A., Lonergan, M., Arthus, M. F., Hendy, G. N., Birnbaumer, M., & Bichet, D. G. (1992). Molecular identification of the gene responsible for congenital nephrogenic diabetes insipidus. *Nature*, 359(6392), 233–235.

Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc*, 5(4), 725–738.

Sadeghi, H., Robertson, G. L., Bichet, D. G., Innamorati, G., & Birnbaumer, M. (1997). Biochemical basis of partial nephrogenic diabetes insipidus phenotypes. *Mol Endocrinol*, 11(12), 1806–1813.

Sadowski, P. G., Groen, A. J., Dupree, P., & Lilley, K. S. (2008). Sub-cellular localization of membrane proteins. *Proteomics*, 8(19), 3991–4011.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4), 406–425.

Seelert, H., Dencher, N. A., & Mller, D. J. (2003). Fourteen protomers compose the oligomer III of the proton-rotor in spinach chloroplast ATP synthase. *J Mol Biol*, 333(2), 337–344.

Sippl, M. J. (1993). Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J Comput Aided Mol Des*, 7(4), 473–501.

Slusarz, M. J., Giedoń, A., Slusarz, R., & Ciarkowski, J. (2006). Analysis of interactions responsible for vasopressin binding to human neurohypophyseal hormone receptors-molecular dynamics study of the activated receptor-vasopressin-g(alpha) systems. *J Pept Sci*, 12(3), 180–189.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol*, 147(1), 195–197.

Sokal, R. & Michener, C. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38, 1409–1438.

Strom, T. M., Hrtnagel, K., Hofmann, S., Gekeler, F., Scharfe, C., Rabl, W., Gerbitz, K. D., & Meitinger, T. (1998). Diabetes insipidus, diabetes mellitus, optic atrophy and deafness (DIDMOAD) caused by mutations in a novel gene (wolframin) coding for a predicted transmembrane protein. *Hum Mol Genet*, 7(13), 2021–2028.

Tan, S., Tan, H. T., & Chung, M. C. M. (2008). Membrane proteins and membrane proteomics. *Proteomics*, 8(19), 3924–3932.

Tanaka, S. & Scheraga, H. A. (1975). Model of protein folding: inclusion of short-, medium-, and long-range interactions. *Proc Natl Acad Sci USA*, 72(10), 3802–3806.

Tanaka, S. & Scheraga, H. A. (1976). Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, 9(6), 945–950.

Tusnady, G. E., Dosztanyi, Z., & Simon, I. (2004). Transmembrane proteins in the Protein Data Bank: identification and classification. *Bioinformatics*, 20(17), 2964–2972.

Tusnady, G. E., Dosztanyi, Z., & Simon, I. (2005). PDBTM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res*, 33(Database issue), D275–D278.

van den Ouweland, A. M., Dreesen, J. C., Verdijk, M., Knoers, N. V., Monnens, L. A., Rocchi, M., & van Oost, B. A. (1992). Mutations in the vasopressin type 2 receptor gene (AVPR2) associated with nephrogenic diabetes insipidus. *Nat Genet*, 2(2), 99–102.

Vargas-Poussou, R., Forestier, L., Dautzenberg, M. D., Niaudet, P., Déchaux, M., & Antignac, C. (1997). Mutations in the vasopressin V2 receptor and aquaporin-2 genes in 12 families with congenital nephrogenic diabetes insipidus. *J Am Soc Nephrol*, 8(12), 1855–1862.

Wettschureck, N. & Offermanns, S. (2005). Mammalian g proteins and their cell type specific functions. *Physiol Rev*, 85(4), 1159–1204.

Willard, L., Ranjan, A., Zhang, H., Monzavi, H., Boyko, R. F., Sykes, B. D., & Wishart, D. S. (2003). VADAR: a web server for quantitative evaluation of protein structure quality. *Nucleic Acids Res*, 31(13), 3316–3319.

Ye, Y. & Godzik, A. (2003). Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, 19 Suppl 2, ii246–ii255.

Zhang, C., Liu, S., Zhou, H., & Zhou, Y. (2004). The dependence of all-atom statistical potentials on structural training database. *Biophys J*, 86(6), 3349–3358.

2.4.2 eProS - A Database and Toolbox for Investigating Protein Sequence-Structure-Function Relationships through Energy Profiles

The eProS database and toolbox as well as graphical front ends have been expanded and improved to enable user-friendly and large-scale energy profile analyses. Annotations from various sources have been retrieved and implemented to provide a broad biological information spectrum. All innovations and improvements form the main part in the discussion of this manuscript which has been submitted for peer-review as a contribution to the annual Nucleic Acids Research database issue.

Published online 31 July 2009

Nucleic Acids Research, 2009, Vol. 37, No. 12 ***1–8***
doi:10.1093/nar/gkn000

ePROS - A Database and Toolbox for Investigating Protein Sequence-Structure-Function Relationships through Energy Profiles

Florian Heinke [1,*], Stefan Schildbach [1], Daniel Stockmann [1] and Dirk Labudde [1,*]

[1]Department of Mathematics, Natural and Computer Sciences, University of Applied Sciences Mittweida, Mittweida, Saxony, Technikumplatz 17, D-09648, Germany

Received January 1, 2009; Revised February 1, 2009; Accepted March 1, 2009

ABSTRACT

Gaining information about structural and functional features of newly identified proteins is often a difficult task. This information is crucial for understanding sequence-structure-function relationships of target proteins and, thus, essential in comprehending the mechanisms and dynamics of the molecular systems of interest.
Using protein energy profiles is a novel approach that can contribute in addressing such problems. An energy profile corresponds to the sequence of energy values which are derived from a coarse-grained energy model. Energy profiles can be computed from protein structures or predicted from sequences. As shown, correspondences and dissimilarities in energy profiles can be applied for investigations of protein mechanics and dynamics. We developed eProS (energy profile suite), a database which provides about 76,000 pre-calculated energy profiles as well as a toolbox for addressing numerous problems of structure biology. Energy profiles can be browsed, visualised, calculated from uploaded structure or predicted from sequence. Furthermore, it is possible to align energy profiles of interest or compare energy profiles with the entire eProS database to identify significantly similar energy profiles and, thus, possibly relevant structural and functional relationships. Additionally, annotations and cross-links from numerous sources provide a broad view of potential biological correspondences. eProS is freely available at http://bioservices.hs-mittweida.de/Epros/.

INTRODUCTION

The amino acid sequence-based prediction of protein structure features, stability analyses of known protein structures as well as secondary structure predictions are important tasks in protein modelling (1). Several energy functions and force fields that model the protein free energy landscape have been developed to address these protein modelling problems. On the one hand, they contribute in protein modelling (i.e. comparative modelling, threading or ab initio folding) and protein model assessment. On the other hand, force fields are essential in molecular simulations and can account for the understanding of dynamics in molecular systems (2). They can also help to comprehend the relations between protein structure and protein function.
Energy models can be based on first principles approaches using physics laws. In addition, statistical analyses of experimentally derived structures lead to the development of so-called knowledge-based energy potentials (KBPs) (2, 3, 4). Although the approaches for computing KBPs are simplified, they can reproduce experimental data with a high level of accuracy if adapted to a specific problem. For example, elastic network models (ENMs) employ simplified coarse-grained interaction models and have proven themselves to accurately determine protein dynamics (5, 6). In general, the continuous application of coarse-grained interaction models is due to the reduction of system complexity and, thus, computational demands.
In 2006, Kozielski and colleagues proposed that the sequences of energy values, so-called energy profiles, derived from protein structures by utilizing potential functions can be compared using modified Needleman-Wunsch (7) and Smith-Waterman (8) alignment procedures. They have shown that pairwise comparisons and detected energy profile similarities can lead to the identification of proteins assigned to the same protein families. Additionally, conformational modifications as results of enzymatic reactions or, in general, protein-environment interactions can be inspected (9, 10, 11). These studies substantiate the possible fields of application of energy profile-based methods. However, to allow large-scale or even data bank-wide investigations, the generation of large datasets is required. Due to their semi-automatic and error-prone nature, generating datasets on a scale comparable to for example the Protein Data Bank (PDB) (12) becomes quite difficult, if physics-based approaches are utilized as proposed by Kozielski et al.
To allow large-scale energy profile-based analyses, we have developed eProS (energy profile suite), a database and toolbox for energy profile-based studying and comparing sequence-structure-function relationships and protein stability. Energy

*To whom correspondence should be addressed. Tel: +49 3727581469; Fax: +49 3727581303; Email: florian.heinke@hs-mittweida.de, dirk.labudde@hs-mittweida.de

profiles are derived by employing a straightforward coarse-grained energy model which is suitable for globular and α-helical membrane protein structures. Currently, eProS stores 74,900 pre-calculated energy profiles derived from experimental globular protein structures and almost 1300 pre-calculated profiles of α-helical membrane protein structures.
The eProS toolbox and the underlying eProS database provide various ways of visualising, downloading and accessing energy profile data. The toolbox also includes database-wide searching for similar energy profiles. Here, the query energy profile can be defined by specifying a structure by PDB-Id, by uploading a structure in PDB format from which the query energy profile is generated, or by uploading an energy profile file which, for example, has been retrieved from the eProS database. Additionally, an amino acid sequence can be queried. Here, an energy profile prediction algorithm is utilized, leading to an energy profile which can be used for database-wide searching. The best matching hits are visualized by the eProS toolbox and/or re-aligned with the query profile. Various sources of annotation (e.g. Gene Ontology (13), PDB, CATH (14), SCOP (15) and Pfam (16)) provide a wide view on structural and functional features of the best hits which can be further broadened via the reverse annotation lookup provided by eProS. The reverse annotation lookup lists all energy profiles that match with the annotation specified by the user. For example, energy profiles of all proteins sharing the same structural topology or molecular function can be investigated concerning common energetic features that point to their similar structure or function. Thus, starting from a protein structure or sequence, estimations about correspondences of protein function and structural features can be drawn from these results and annotations.

DATABASE AND SEARCH TOOL DESCRIPTION

Content and data organization.

At present eProS supplies energy profiles for approximately 76,200 PDB entries which are internally separated into 74,900 globular protein and 1,300 α-helical transmembrane protein energy profiles. The corresponding PDB flat files, containing protein structure information, are stored in a local directory on the web server hosting eProS. Based on these files energy profiles for each available PDB entry have been pre-computed. Energy profiles are stored in files of a specifically tailored format. The following file formats have been defined and are available for download and analyses:

- *.ep: Tab character separated files with each row containing five columns (chain identifier, PDB residue index, amino acid one-letter code, secondary structure assignment and energy value). The first two lines are reserved for listing the PDB-Id and the header row.
- *.ep2: Extended *.ep file. Each line represents a record, whereat the first four characters specify the record type. Record fields are tab character-separated. The following record types are currently defined: "NAME" (PDB-Id), "TYPE" ("TM" for α-helical transmembrane protein structures, "nTM" for globular protein structures), "HEAD" (header row), "ENGY" (energy value for a single residue, the five record fields correspond to the five columns of a row in a *.ep file) and "REMK" (indicating a comment line).
- *.eps: Binary files. This file format is going to be used in upcoming standalone software applications. It is only in use in server-internal routines at present.

For each protein structure in the local PDB file repository an energy profile have been saved in each of these three file formats.

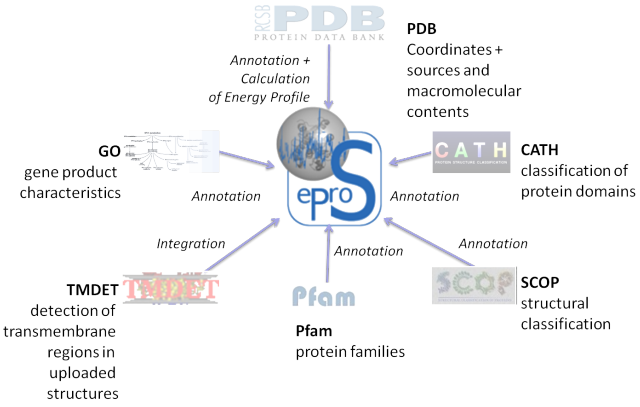


Figure 1. The eProS database integrates data from the Gene Ontology database (13), PDB (12), CATH (14), SCOP (15) and Pfam (16). Energetic discrepancies and similarities between proteins can be investigated by the eProS toolbox. The energy profile annotations provided by the data integration can broaden the understanding of the relationships between energetic and functional and structural properties. Furthermore, the advanced reverse annotation lookup can be a valuable method to identify functional and structural related proteins and study their energetic similarities.

The annotation entries displayed on the detailed view of a protein are retrieved from internal relational databases at runtime. All Information provided by the databases have been obtained from external sources. An overview of the related databases from which data has been obtained is given in Figure 1. Detailed data integration has been achieved as follows:

TMDET prediction: TMDET is a service which implements a neural network-based method for the prediction of membrane spanning regions in three-dimensional structures (17). For each α-helical membrane protein structure present at eProS database a prediction have been computed. These predictions are essential for deriving an energy profile from α-helical membrane protein structures by employing the coarse-grained model (for details, see). Additionally, if the user has uploaded a α-helical membrane protein structure for analyses, TMDET is applied for prediction and, from this the energy profile is generated.

Pfam classification: The current release of the Pfam database is available in terms of an SQL dump (16). As for the annotation retrieval, only two tables are required: 'pfama' (approx. 2,200 rows), containing the Pfam classifications and 'pdb_pfama_req' (approx. 110,000 rows), mapping PDB-IDs to their corresponding classifications.

SCOP classification: The SCOP classification releases are not provided in terms of an SQL database dump but as character separated value files instead (15), from which

the following tables resulted: 'des' (approx 144,000 rows) containing descriptions for all SCOP classifications, 'hie' (approx. 144,000 rows) representing the SCOP classification hierarchy and 'cla' (approx. 111,000 rows) assigning PDB-Ids to SCOP classifications.

CATH classification: Two files of the current CATH database release (14) served as source of annotation data: 'CathDomainList' and 'CathNames'. The resulting tables 'domains' (approx. 153,000 rows) and 'names' (approx. 3,900 rows) provide information about assigned CATH domains and names of CATH classification nodes respectively.

GO term annotation: Out of the 44 tables found in the images of the GO database (13) the following are required to gain the GO terms associated with a protein structure: 'term' (approx. 37,000 rows) containing the GO terms, 'gene_product' (approx. 13,000,000 rows) containing gene products, 'species' (approx. 890,000 rows) containing species the gene products originate from. The 'association' table lists (approx. 77,000,000 rows) assignments of GO terms to gene products. In order to find GO terms matching to a PDB-Id, the name of each macromolecule of the entry and their sources (NCBI Tax ID) are required. Therefore, two auxiliary tables have been created (approx. 122,000 rows and 108,000 rows respectively). However, searching for GO terms and especially performing annotation reverse lookups had caused unacceptable response times due to complex query statements during the development. For performance improvements, an additional table (approx. 610,000 rows) has been created, which assigns the GO terms to each protein directly. Thereby a speedup ($>$ 5 min. to 100 ms) has been achieved for reverse annotation lookup.

Working with the eProS database

The eProS and the collection of energy profiles are freely available to the scientific community in a separate download section (accessible via the 'Dataset' link at the eProS homepage). In the download section it is possible to browse the database and inspect energy profiles of interest. For this purpose, eProS provides the access of energy profile data via flat HTML page tables or by automated download programs, such as wget and similar software. This ensures large-scale downloading of energy profile files and high-throughput analyses. In contrast, the eProS toolbox permits accessing the data by more sophisticated energy profile visualisations, e.g. plotting of energy profiles and viewing the protein structure of interest with energy value-based coloring schemes. Cross-links and annotations retrieved from various foreign sources (see Figure 1) are available for the user. From these annotations, the reverse annotation lookup can be accessed and, subsequently, energy profiles of proteins matching the user-specified annotation are listed. As an example, after querying the N-terminal domain of the riboflavin synthase by specifying its PDB-Id (1pkv) at the eProS home page (see Figure 4A), the corresponding energy data and structure as well as the related annotations are listed (Figure 4B). Reverse annotation lookup is accessed by clicking the annotation of choice, which leads to the list of energy profile data available at eProS that share the specified annotation, in this case riboflavin synthase activity (Figure 4C).

Further methods have been implemented and integrated into the toolbox that allow energy profile analysis, calculation, sequence-based prediction and a database-wide searching for identical or similar energy profiles. An overview of these tools and the implemented data flow is given in Figure 2. The following elucidations explain these tools briefly:

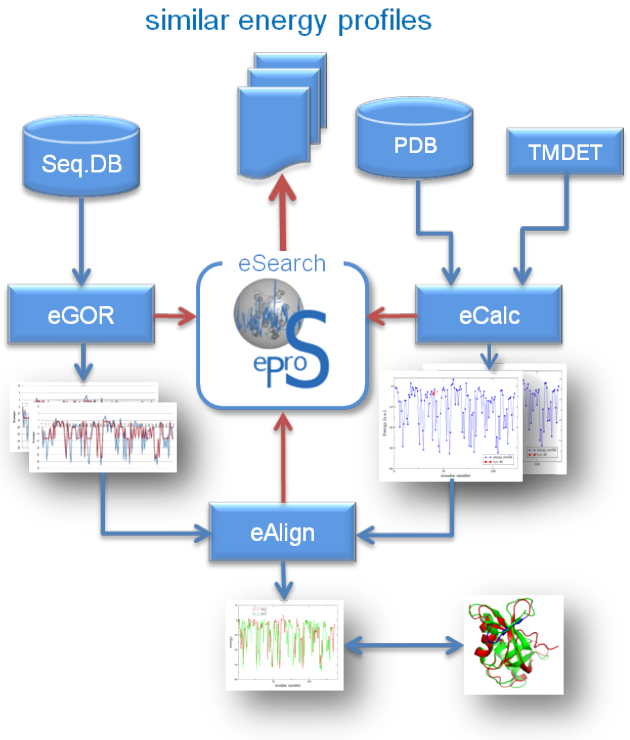


Figure 2. The eProS toolbox affords several, interconnected methods for accessing and working with the eProS database. Besides energy profile data access and calculation (eCalc), methods for predicting (eGOR) and aligning (eAlign) energy profiles have been implemented. These techniques can be utilized to derive or retrieve energy data which can be queried to the eProS database (eSearch). Identified significantly similar energy profiles can be further investigated. Provided annotations and detected similarities can aid in understanding the dynamics and functional aspects of the protein of interest.

eAlign: This tool provides modified Needleman-Wunsch (7) and Smith-Waterman-like (8) alignment procedures for computing pairwise energy profile alignments. Generated alignments are presented as graphs, ASCII-formatted texts as well as dotplots in which energetic identities and similarities are highlighted. In addition, eAlign computes a so-called distance Score (dScore). The dScore gives a hint about the energy profile similarity observed in the alignment.

eCalc: eCalc provides the energy profile computation as described in the section. On the one hand, a PDB-Id can be specified and the corresponding energy profile data is displayed if present in the database. If the entry is not present, the coordinates are retrieved from the PDB and the energy profile is computed. Additionally, the user can upload a protein structure from which the energy profile will be generated. This data can be investigated in at the eProS output site, downloaded for further analyses or reused as input for the eProS toolbox.

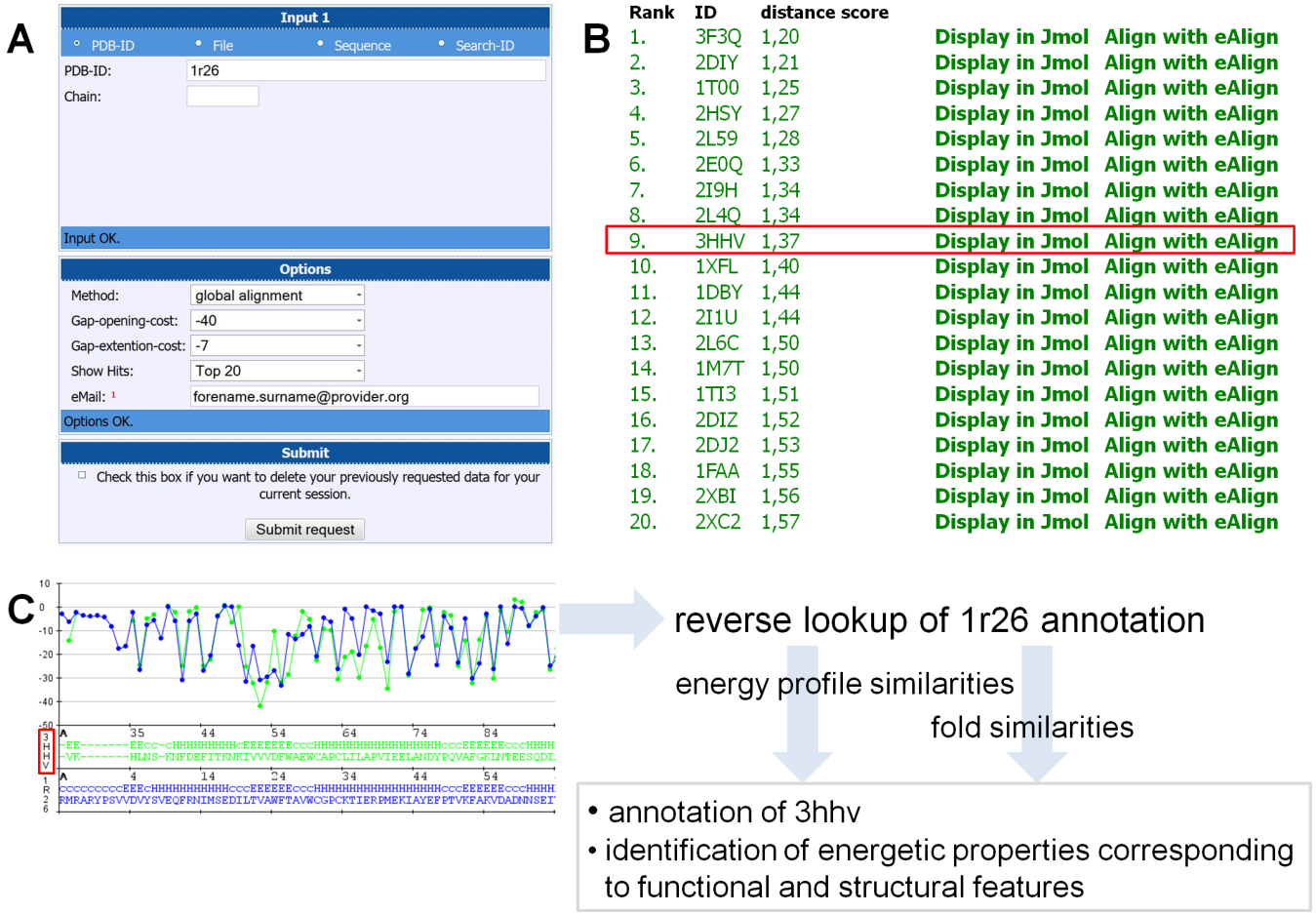


Figure 3. eSearch is a tool that provides eProS-wide searching for similar protein energy profiles. It can be queried (A) by specifying an energy profile by its PDB-Id, by uploading a structure from which the corresponding energy profile is subsequently computed or by specifying a sequence. In the last case, an energy profile prediction algorithm (eGOR) is applied. Because of long computing times, the user has to enter a valid e-mail address. Access to previous search results is provided by direct-links and session ids send by e-mail. eSearch results are given as lists which are ranked according to derived dScores (B). The dScore reports the significance of the resulting energy profile alignment and, thus, energy profile similarity. In this example, the *Trypanosoma brucei brucei* thioredoxin has been queried, leading to a list of various energetically similar protein structures (hits). For the ninth hit (PDB-Id 3hhv), no GO-term annotation is given. With respect to the energy profile alignment (C) and similarities obtained from energy profile analyses of further hits and entries retrieved by reverse annotation lookup, functional and structural energetic properties can be detected which are common in all identified energy profiles, including 3hhv. From this, GO-term annotations for 3hhv can be proposed.

eGOR: A modified GOR algorithm (eGOR) have been developed which allows energy profile prediction from sequence. This algorithm is the basis for the eGOR tool and allows the prediction of an energy profile from a user-specified sequence.

eMut: This tool visualises the energetic similarities and dissimilarities of proteins with the same length. Thus, analysis of, for example, point-mutated proteins, molecular dynamics or coarse-grained dynamics trajectories (i.e. trajectories generated from ANM) or influences of different temperatures on protein stability.

eSearch: The eSearch tool facilitates database-wide searching for identifying similar energy profiles to a user query. The query energy profile can be specified in various ways. First of all, eSearch enables searching by a user-specified PDB-Id. Additionally, the user can provide a protein structure (f.e. a structure model) by uploading the coordinate data in PDB format from which the energy profile is generated and queried to the eProS database. Third, an energy profile file can be uploaded which has been generated by means of eCalc, eGOR or which has been retrieved from the database.

In the process, pairwise alignments of the query energy profile to all entries of the specified entry set (e.g. globular proteins or α-helical membrane proteins) are generated. From each alignment, the corresponding dScore is heuristically computed and recorded. This process requires about three hours of computation after querying an average-sized protein structure ($\approx$ 120 aa) to the set of globular protein energy profiles. Because of the time limitation, the user has to specify a valid e-mail address in order to run an eSearch query. After the computation has finished, an email is send to the user which provides a link to the result session as well as a session id. The results are presented as an interactive list ranked according to the derived dScores. An example of using eSearch is illustrated in Figure 3. After querying the energy profile of *Trypanosoma brucei brucei* thioredoxin (PDB-Id 1r26), numerous similar energy profiles are identified (see

Figure 3A and B). As a representative example for the general observations that can be made from this query, the energy profile alignment to the ninth match (PDB-Id 3hhv) indicates numerous global energetic correspondences (Figure 3C). As shown, the best matching energy values are located in the first helix and second strand in both structures. By utilizing the reverse annotation lookup of functional annotations (e.g. 'cell redox homeostasis' and 'glycerol ether metabolic process') and structural annotations (e.g. 'glutaredoxin') of 1r26, corresponding energy profile entries are listed. Note, that most proteins present in this list are reported as best-matching energy profiles. Since 3hhv has not been annotated by GO-terms yet, it can be proposed that the GO-terms associated to the best matches can be applied for annotating 3hhv. Furthermore, integrating the profiles reported by eSearch and reverse annotation lookup to the analyses, the energetic properties can be identified that are responsible for stabilizing the fold. For example, mainly low-energetic residues can be found in the second strand. In contrast, the third strand is consisting of residues with alternating energy values. Both observations are in agreement in all proteins sharing this topology. On the other hand functional energetic features might be basically corresponding to residues located in the first helix and second strand, since these residues are found to be energetically conserved in all energy profiles listed by the functional reverse annotation lookup.
In a similar way, the functional clarification of protein structures of unknown function can be facilitated.

THEORY OF PROTEIN ENERGY PROFILES

The coarse-grained energy model applied for protein energy profile-based analyses available at eProS belongs to the KBPs, whereby energy computation is derived from statistics and concepts of statistical physics. In essence, the energy of a residue is approximated according to the amino acid buriedness propensity. The buriedness propensity of each of the 20 amino acids o is derived from the number of occurrences to find o being exposed at the protein surface ($n_{o,out}$) or being buried in the protein structure ($n_{o,in}$). Exposed/buried states can be determined by employing the following equation:

$$\left\|C_{\alpha,o}-c\right\|_{D_E}<5\text{Å}\vee\langle C_{\beta,o}-C_{\alpha,o},c-C_{\alpha,o}\rangle>0 \qquad (1)$$

Here, c corresponds to the Cartesian center of all $C_{\alpha,r}$ atoms with $\left\|C_{\alpha,o}-C_{\alpha,r}\right\|_{D_E}<10\text{Å}$. We used a set of 2,700 non-redundant globular and 380 non-redundant α-helical membrane protein structures to derive the buriedness propensities for each protein type. In α-helical membrane proteins, the topology of the structure needs to be taken into account, since amino acid propensities differ significantly between regions which are located inside the hydrophobic membrane environment and regions which are located at the extra/intracellular side of the membrane. Thus, energy calculation of a residue requires a further inside/outside assignment s which is relative to the membrane bilayer. For this purpose, TMDET (17) is applied in our work and which is implemented at eProS as well. By taking the topology assignment s into account, the energy of a residue i can be computed according to the inverse Boltzmann principle as follows:

$$e_i=-k_BT\ln\left(\frac{n_{i,in,s}}{n_{i,out,s}}\right)+k_i\ , \qquad (2)$$

with $k_i=0$, if i is located at the extra/intra-cellular site of the membrane ($s=$nTM), or

$$k_i=-k_BT\ln\left(\frac{n_{i,\text{TM}}}{n_{i,\text{nTM}}}\right) \qquad (3)$$

otherwise. In the process, the temperature T is declared as a constant, whereby k_BT can be neglected from the calculation and resulting energy values are given as arbitrary units. Note, in the case of investigating a globular protein, propensities derived from globular protein structures need to be applied. In this case, the energy model has been more simplified since no topology needs to be considered:

$$e_i=-\ln\left(\frac{n_{i,in}}{n_{i,out}}\right). \qquad (4)$$

To approximate the total energy E_i^* all energies of all interacting residues j are taken into account (18):

$$E_i^*=\sum_j f(i,j)\left(e_i+e_j\right)\ , \qquad (5)$$

with

$$f(i,j)=\begin{cases}1, & \|i-j\|_{D_E}<8\text{Å}\\ 0, & \text{else.}\end{cases} \qquad (6)$$

In the process C_β or, in case of observing glycine, C_α atom coordinates are declared as spatial representatives. The sequence of all E_i^* of a given protein structure corresponds to the protein energy profile. Since the energy profile calculation integrates spatial and physico-chemical information, an energy profile can be interpreted as a protein-specific representation. As an example, the eProS output of the human angiogenin variant H13A (PDB-Id 1b1j) is illustrated in Figure 5. The sequential visualisation of an energy profile (see Figure 5B and D) is the most intuitive. However, energy-to-structure mappings, as shown in 5C, can contribute to identify low-energetic and, thus, stabilizing regions and properties in protein structures.
Similar to the methodology introduced by Kozielski and colleagues, methods for computing pairwise energy profile alignments have been implemented. By that, energetic shifts can be analysed and global energy profile distances can be derived. From each energy profile alignment a so-called dScore is derived as a measure of similarity significance. Alignments of identical energy profiles correspond to a dScore of 0. In contrast, the maximum distance derived by this approach (e.g. a dScore of 5) indicates no detectable similarity. The observation of the dScore is the basis for user-friendly searching in the eProS databases.

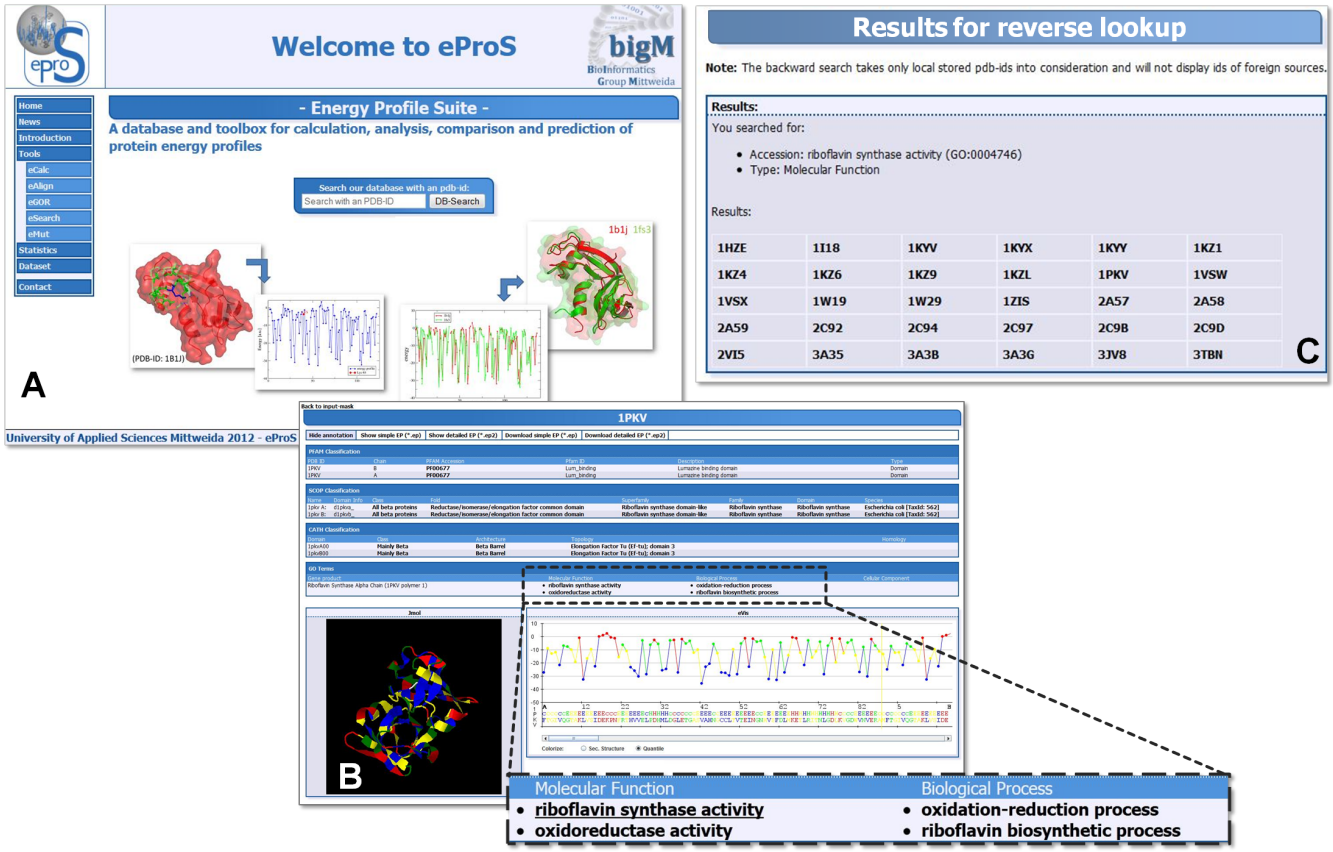


Figure 4. The eProS database can be accessed via the eProS homepage (A) by specifying a PDB-Id, by browsing the dataset ('Dataset' link on the left in A) or by utilizing the eCalc tool of the eProS toolbox (see list in A on the left). The graphical output of energy profile data given by eProS (B) includes visualisations by graphs and a 3D structure viewer. Annotations related to foreign databases (see Figure 1) are reported and provide a wide spectrum of information about structural and functional aspects of the protein of interest. Additionally, annotations can be accessed interactively. By that, the integrated reverse annotation lookup lists all PDB-Ids which match the specified annotation and which are currently present at the database. From this, the listed protein structures can be investigated concerning common energetic properties that point to their structural and functional similarity.

Correspondences of energy, functional characterization by GO-terms and structure

To investigate correlations between coarse-grained energies and energies derived by molecular dynamics (MD), 220 globular protein structures were obtained from the PDB and analysed by the TINKER molecular dynamics software suite (19). By this analysis, all-atom energies were computed and investigated for correlations to energies derived by the coarse-grained energy model. As shown in Figure 6A, total binding energies calculated by TINKER (E_{FG}) correlate well to the sums of all energies computed by the coarse-grained energy model (E_{CG}). Thus, this coarse-grained energy model can be used to draw physical and biological meaningful conclusions concerning residue stability as well as destabilizing effects and alternations resulting from various causes, i.e. point mutations, deletions, conformational rearrangements or protein-protein interactions.

Furthermore, sequential, structural and functional correspondences to pairwise energy profile distances have been investigated. For this purpose, 2,700 non-redundant globular protein structures and their corresponding GO-term annotations have been obtained from eProS. Sequence identities and structural similarities have been recorded as well. To investigate functional correspondences, GO-term annotations have been compared semantically using the G-SESAME web server (20). As depicted in Figure 6B, sequence identities (seqId), structural similarities ($-\log$(p-value), calculated by FATCAT (21), function similarity (semantic GO-term annotation similarity, illustrated by a blue-to-red coloring scheme) correlate to dScores. This substantiates that energy profiles are abstractions of sequential, structural and functional information. Additionally, these correspondences can be transferred to α-helical membrane proteins as well. Thus, energy profile differences correspond to functional and structural divergences and can be analysed in detail. Global or local energy profiles alignments methods provide the identification of residues which are alternating energetically and which might be the triggers of the observed molecular effects.

CONCLUSION

The analyses of proteins based on energy profiles can contribute in understanding correspondences of protein sequence to structure, stability and function (9, 10, 11, 18). However, the generation of large scale databases of energy profiles derived from physics-based approaches are difficult to automatise, error-prone and slow. The eProS database and the eProS toolbox provide energy profile-based calculation, analysis,

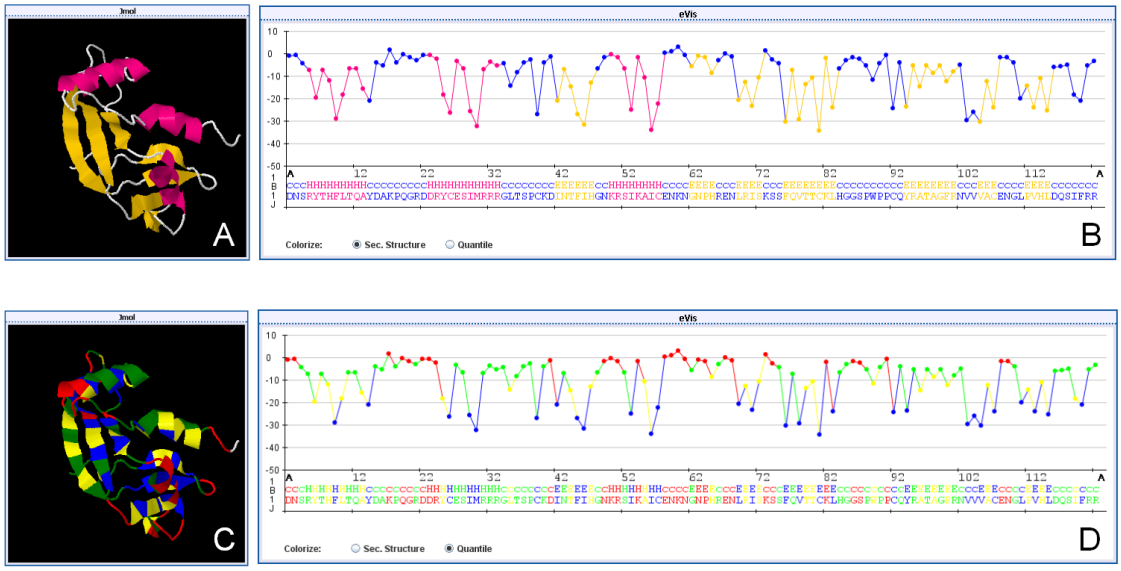


Figure 5. Protein energy profiles are derived from integrating spatial and physico-chemical information and can be interpreted as a protein-specific representation. In this figure, the eProS output of the human angiogenin variant H13A (PDB-Id 1b1j) is depicted. In B and D, the energy profile is shown as a sequence of energy values. Coloring schemes for energy-to-structure mappings (C) and structure-to-sequence mappings (B) are provided. The energy coloring is discretised by assigning each energy value to one of the 4-quantiles in the energetic spectrum derived from the eProS database. This measure is due to visualisation and performance purposes. The color mappings provide insights which can contribute to identify low-energetic and, thus, stabilizing regions and properties in protein structures.

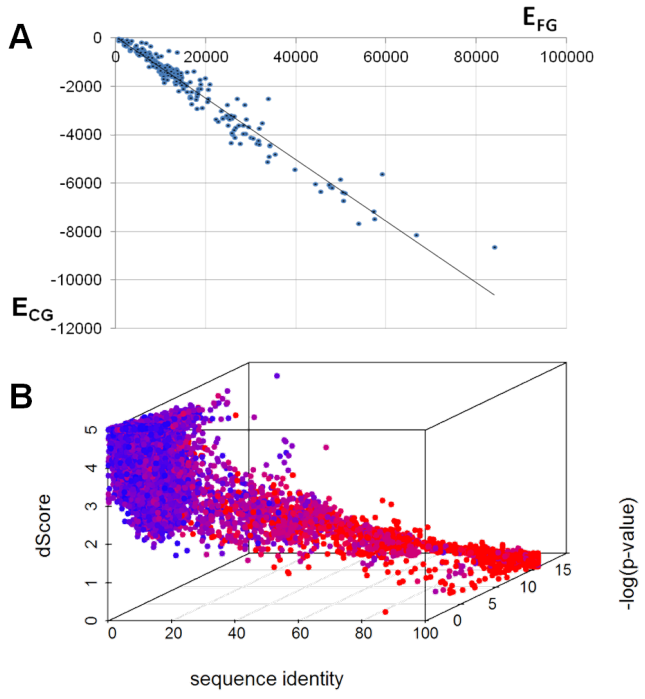


Figure 6. dScores of pairwise alignments of 2,700 protein structures correlate well to derived sequence identities and structure similarities (indicated by p-values on −log-scale) as well as functional correspondences (indicated by blue-to-red-coloring scheme with blue-coloring highlighting alignments of proteins with no functional correspondences). This correlation substantiates the applicability of energy profiles for investigating relationships of protein sequence, structure and function as well as analysing functional alterations as effects of mutations or protein-environment interactions.

prediction and comparison of protein energy profiles computed by a coarse-grained energy model. Cross-links and annotations which are related to structure and annotation databases provide a wide information spectrum of the protein of interest. eProS also provides reverse lookup techniques for browsing energy profiles that match a user-specified annotation of the investigated protein. From this, energetic correspondences can be identified which might correspond to structural and functional features. Such insights can be helpful for coarse-grained analyses of protein dynamics and protein-protein interactions. Additionally, the investigation of protein families on the basis of energy profiles can contribute to elucidate family memberships and functional variability. Especially the eGOR algorithm provided by the eProS toolbox can aid in approaching such biological questions.

Future developments of eProS are going to include the improvement of time-performance of eSearch. Additionally, implementing cross-links between energy profile data will provide a more user-friendly data access. Furthermore, an automated energy profile and annotation retrieval system is currently work in progress which is capable of updating the eProS database on a weekly or monthly basis. At the moment, an approach for predicting the topology of a α-helical membrane protein based on its predicted energy profile is currently under evaluation and will be integrated to the toolbox.

ACKNOWLEDGEMENTS

The authors would like to thank the group members Steffen Grunert and Michael Spranger. We acknowledge funding by "Europäischer Sozialfonds" (ESF), the Free State of Saxony and the University of Applied Sciences Mittweida.

Conflict of interest statement. None declared.

REFERENCES

1. M. Zvelebil and J.O. Baum. *Understanding Bioinformatics*. Garland Science, 2008.
2. Chi Zhang, Song Liu, Hongyi Zhou, and Yaoqi Zhou. The dependence of all-atom statistical potentials on structural training database. *Biophys J*, **86**(6):3349–3358, Jun 2004.
3. M. J. Sippl. Calculation of conformational ensembles from potentials of mean force. an approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol*, **213**(4):859–883, Jun 1990.
4. M. J. Sippl. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J Comput Aided Mol Des*, **7**(4):473–501, Aug 1993.
5. A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys J*, **80**(1):505–515, Jan 2001.
6. Andrzej Kloczkowski, Robert L. Jernigan, Zhijun Wu, Guang Song, Lei Yang, Andrzej Kolinski, and Piotr Pokarowski. Distance matrix-based approach to protein structure prediction. *J Struct Funct Genomics*, **10**(1):67–81, Mar 2009.
7. S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, **48**(3):443–453, Mar 1970.
8. T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, **147**(1):195–197, Mar 1981.
9. D. Mrozek, B. Malysiak, and S. Kozielski. EAST: Energy Alignment Search Tool. In Lipo Wang, Licheng Jiao, Guanming Shi, Xue Li, and Jing Liu, editors, *Fuzzy Systems and Knowledge Discovery*, volume **4223** of *Lecture Notes in Computer Science*, pages 696–705. Springer Berlin / Heidelberg, 2006.
10. D. Mrozek, B. Malysiak, and S. Kozielski. An optimal alignment of proteins energy characteristics with crisp and fuzzy similarity awards. In *FUZZ-IEEE'07*, pages 1–6, 2007.
11. D. Mrozek, B. Malysiak-Mrozek, and S. Kozielski. Alignment of protein structure energy patterns represented as sequences of fuzzy numbers. In *Fuzzy Information Processing Society, 2009. NAFIPS 2009. Annual Meeting of the North American*, 2009.
12. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res*, **28**(1):235–242, Jan 2000.
13. M. Ashburner et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**(1):25–29, May 2000.
14. Alison L. Cuff, Ian Sillitoe, Tony Lewis, Andrew B. Clegg, Robert Rentzsch, Nicholas Furnham, Marialuisa Pellegrini-Calace, David Jones, Janet Thornton, and Christine A. Orengo. Extending cath: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Res*, **39**(Database issue):D420–D426, Jan 2011.
15. A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, **247**(4):536–540, Apr 1995.
16. M. Punta et al. The Pfam protein families database. *Nucleic Acids Res*, **40**(Database issue):D290–D301, Jan 2012.
17. Gábor E. Tusnády, Zsuzsanna Dosztányi, and István Simon. TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics*, **21**(7):1276–1277, Apr 2005.
18. F. Heinke and D. Labudde. Membrane protein stability analyses by means of protein energy profiles in case of nephrogenic diabetes insipidus. *Comput Math Methods Med*, 2012:790281, 2012.
19. J. Ponder. TINKER – software tools for molecular design. Technical report, Dept. of Biochemistry and Molecular Biophysics, Washington University, School of Medicine, St. Louis, 2001.
20. Z. Du, L. Li, C. F. Chen, P. S. Yu, and J. Z. Wang. G-SESAME: web tools for GO-term-based gene similarity analysis and knowledge discovery. *Nucleic Acids Res.*, **37**:W345–349, 2009.
21. Yuzhen Ye and Adam Godzik. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, **19** Suppl 2:ii246–ii255, Oct 2003.

2.4.3 Structure topology prediction of discriminative sequence motifs in membrane proteins with domains of unknown functions

This manuscript has been submitted for peer-review to ISRN Bioinformatics. As the second author I contributed by proposing algorithmic concepts and data visualizations. This study emphasizes the identification and characterization of short sequence motifs widely present in membrane protein families. The elucidated analyses lead to an approach suitable for predicting topological states of each motif. However, certain motifs show insufficient prediction accuracy over all investigated protein families. We propose that these short sequence motifs can be separated into basic structure-forming elements on the one hand, serving as essential structural modules mediating spatial interactions which might be crucial for membrane protein folding. On the other hand, motifs showing insufficient prediction accuracies might correspond to motifs that are important in defining protein function. The application of energy profile-based methods can be applied to inspect energetic properties of the investigated motifs. From this, the characteristics can be derived that determine topological states or protein function.

ISRN Bioinformatics

Pages 1–7

Structure topology prediction of discriminative sequence motifs in membrane proteins with domains of unknown functions

Steffen Grunert[1,*], Florian Heinke[1,*] and Dirk Labudde[1]

[1]Hochschule Mittweida, University of Applied Sciences, Germany, Technikumplatz 17, 09648 Mittweida

ABSTRACT

Motivation: Membrane proteins play essential roles in cellular processes of organisms. Photosynthesis, transport of ions and small molecules, signal transduction and light harvesting are examples of processes which are realized by membrane proteins and contribute to a cell's specificity and functionality. The analysis of membrane proteins has shown to be an important part in the understanding of complex biological processes in the context of proteomics and genomics. Genome-wide investigations of membrane proteins have revealed a large number of short, distinct sequence motifs. These motifs support the understanding of the features that are important for establishing stability and functionality of the folded membrane protein in the membrane environment. Thus, membrane protein sequence motif analysis can be helpful in a number of applications, e.g. the investigation of mutant proteins and potential effects of mutagens.

Results: The analysis of 32 membrane protein families with domains of unknown functions (DUF) discussed in this study led to a novel approach which describes the separation of motifs by residue-specific distributions. Based on these distributions we can predict the topology of the majority motifs in putative membrane proteins with unknown topology and function.

Conclusion: We hypothesize that short sequence motifs can be separated into structure-forming motifs on the one hand, as such motifs show high prediction accuracy in all investigated protein families. This points to their general importance in α-helical membrane protein structure formation and interaction mediation. On the other hand, motifs which show high prediction accuracies only in certain families can be classified as functionally important and relevant for family-specific functional characteristics.

Contact: sgrunert@hs-mittweida.de

1 INTRODUCTION

Membrane proteins are essential for many fundamental biological processes within organisms. Active nutrient transport, signal and energy transduction or ion flow are only a few of the numerous functions enabled by membrane proteins [1]. Membrane proteins obtain their specific functionality by individual folding and interactions with the hydrophobic membrane environment as well as, in many cases, by oligomeric complex formation and protein-protein interactions [1; 2]. The identification of such complexes and interactions is valuable, since, on the one hand, detailed information of the function of an unknown membrane protein can be obtained by analysing its interactions with proteins of known function. On the other hand, biological processes can be comprehended as a dynamically fluctuating system, whereby the biological role of the unknown membrane protein can be defined more precisely [3; 4]. Accordingly, destabilization of the three-dimensional structure of a membrane protein caused by mutations or ligand interactions are triggers for numerous diseases, f. e. diabetes insipidus, cystic fibrosis, hereditary deafness and retinitis pigmentosa [5; 6; 7].

Although 20–30% of all open reading frames of a typical genome are encoding membrane proteins [5; 8; 9] and 60% of all drug targets are membrane proteins [2]. Membrane proteomics is still an experimentally challenging field due to poor protein solubility, wide intra-cellular concentration range and, thus, inaccessibility to many proteomics methodologies [10]. Hence, the number of known three-dimensional structures is relatively small, with 394 non-redundant membrane protein chains currently available [11; 12; 13]. Therefore, there is a necessity for approaches that allow to predict structural and functional features of unknown membrane proteins. A variety of methods have been developed to predict structural features from sequence, such as α-helical membrane-spanning helices and extra/intra-cellular domains (i.e. TMHMM [14], PHDhtm [15], MEMSAT3 [16]) as well as membrane-spanning beta strands of transmembrane β-barrel proteins (i.e. BOCTOPUS [17]). Furthermore, in genome-wide membrane protein sequence analyses, numerous short conserved sequence motifs were identified [18]. As an example, the most widely discussed GxxxG motif has been shown to be significantly present in transmembrane α-helices. With both glycines resting on one side of the helix as spatially neighbouring residues and by that forming a smooth helix membrane surface, structural studies confirmed that the GxxxG motif plays an important part in mediating helix-helix interactions [18; 19; 20; 21; 22]. In general, short conserved membrane protein motifs are considered to be significantly relevant for membrane protein folding and structural stability as well as being involved in defining a protein's function.

*to whom correspondence should be addressed

Hence, sequence motif analyses and resulting insights can support the understanding of protein dynamics. Information can be derived which may contribute to study the dynamics of mutant proteins and the effects of mutagens [23; 24; 25]. Additionally, as addressed in [26], the analysis of sequence motifs in proteins with similar function or structure might help to identify essential functional sites and locations which contribute to structural stability.
In this work we focused on previous studies and results that have been reported by Gerstein and colleagues [18]. In the process, various integral membrane protein families with polytopic membrane domains had been obtained from Pfam database [27]. As part of their studies, locations of the at least conserved residues (glycine, proline and tyrosine) in α-helical transmembrane regions had been investigated. As a result, short motifs consisting of pairs of small residues (glycine, alaine and serine) surrounding single or multiple variable positions had been identified in conserved sequences and Pfam-classified families. Based on these results we have developed a prediction approach to allocate the topological state of a sequence motif in the protein structure based on sequence information. We have used cross-validation to verify the prediction accuracy. However, prediction accuracy has been found to be variable for certain motifs with regard to the investigated protein family. According to this, we hypothesize that short sequence motifs can be separated into structure-forming motifs on the one hand, as such motifs show high prediction accuracy in all investigated protein families. This points to their general importance in α-helical membrane protein structure formation and interaction mediation. On the other hand, motifs which show high prediction accuracies only in certain families can be classified as functionally important and relevant for family-specific functional characteristics.

2 MATERIALS AND METHODS

2.1 Used membrane protein families

As the first step of our analysis 32 membrane protein families with domains of unknown functions (DUF) were obtained from the Pfam database [27] using extended keyword searching. All 7051 sequences were retrieved for statistical analysis. The full list of employed membrane protein families is given in Table 2.1. Subsequently, 50 sequence motifs, identified by Gerstein and colleagues [18], were localized in the obtained set of families.

2.2 Programs and tools

To avoid generating misguiding statistics by including identical or highly similar sequences, a set of non-redundant sequences was generated. Here, we defined the sequence redundancy threshold at 25% sequence identity. In the first step of sequence processing, CD-HIT [29] was applied for first clustering. However, CD-HIT accepts only non-redundancy thresholds of $>$40%. This limitation is caused by the internal word-length filtering approach and statistical presets. Hence, to ensure clustering sensitivity, a 60% non-redundancy threshold, which corresponds to tetrapeptide word filtering used by the program, was applied. In the second step, sequence clusterings using the 25% redundancy threshold were obtained by means of utilizing BLASTClust [30]. The representative sequences of all cluster were extracted, leading to a set of 2511 non-redundant sequences.

Subsequently, the determination of membrane and non-membrane associated sequence regions was computed using TMHMM Server v. 2.0 [14].

Basically, TMHMM performs a prediction of intra/extra-cellular regions and integral membrane helices starting from sequence. Additionally, the probability of the prediction is given for each residue, as well. According to the results we obtained from TMHMM, a topological state was assigned to each residue. A residue was assigned as 'TM' if the posterior prediction probability of this residue being a part of a membrane helix has been found to be greater than 90%. If the posterior prediction probability of the residue has been greater 90% for extra/intra-cellular prediction, the residue was assigned as 'nTM'.

Table 1. 32 membrane protein families were derived from Pfam database[28] and employed for statistical analysis.

Accession	Family	Accession	Family
PF09767	DUF2053	PF09945	DUF2177
PF09834	DUF2061	PF09946	DUF2178
PF09842	DUF2069	PF09971	DUF2206
PF09843	DUF2070	PF09972	DUF2207
PF09852	DUF2079	PF09973	DUF2208
PF09858	DUF2085	PF09980	DUF2214
PF09874	DUF2101	PF09990	DUF2231
PF09877	DUF2104	PF09991	DUF2232
PF09878	DUF2105	PF09997	DUF2238
PF09879	DUF2106	PF10002	DUF2243
PF09880	DUF2107	PF10011	DUF2254
PF09881	DUF2108	PF10067	DUF2306
PF09882	DUF2109	PF10080	DUF2318
PF09900	DUF2127	PF10081	DUF2319
PF09913	DUF2142	PF10097	DUF2335
PF09925	DUF2157	PF10101	DUF2339

2.3 Used motifs

The short sequence motifs analysed in our work have been reported in [18]. In this study, Gerstein et al analysed consensus sequences of 168 Pfam-A families to identify significant amino acid pair motifs. By the comparison of their results in earlier published findings (see [20]), a list of 50 significant motifs has been derived which we used in our work (for original data see [18], Table 1, List 3): GG4, GL3, GG7, GL1, AG7, GA7, AG4, PL2, AS4, AL6, LP1, PG9, GA4, FG1, SL1, SG4, PL1, AA7, AG5, LF8, IA1, GV1, AI1, AA2, GL2, AA3, SL2, PG5, PG6, IL4, GS5, VL4, GV2, IG1, PG10, LY6, LF10, SA6, LG5, SA3, PF1, GS4, IV4, LS1, GY8, IG2, LF9, VF8, VG6, GN4

Intuitively, the reported short sequence motifs can be written in a generalized, regular expression-like form of XYn, where X and Y correspond to amino acids separated by n-1 highly variable positions. However, in the process of analysis we found that short motives with a relatively small number of variable positions (more precisely, if n is found to be < 3) don't contain enough information to be investigated by our approach. Thus, these motifs have been discarded in the process, which resulted to a final set of 33 sequence motifs. In our non-redundant sequence set, almost 250,000 single motif occurrences were identified. As an example of motifs located in a membrane protein structure, Figure 1 illustrates the seven motif occurrences which can be found in the bacteriorhodopsin trimer (PDB-Id: 1brr).

2.4 Information extraction and clustering

In this work we will show a novel approach is elucidated which predicts the topology state of a short sequence motif occurrence in membrane proteins. The following steps were completed to establish this approach.
At first, all single motif occurrences were identified in the non-redundant

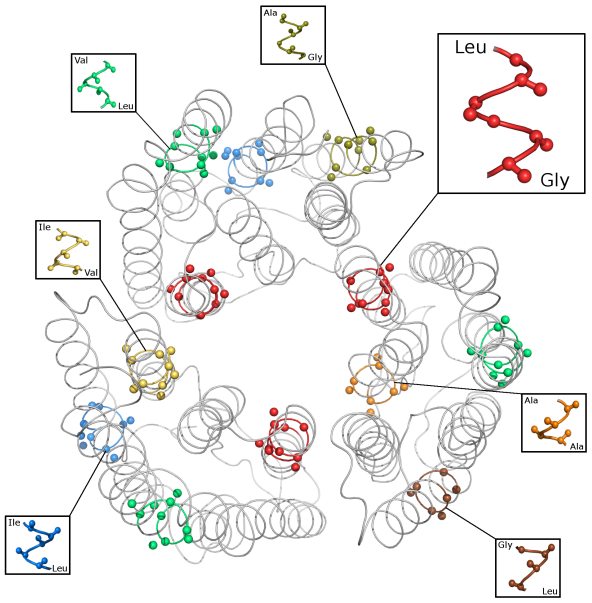


Figure 1. In the bacteriorhodopsin trimer (PDB-Id: 1brr), seven of 33 sequence motifs, which were analysed in this study, are present. Each motif can be written in a regular-expression like of XYn, where X and Y are amino acids separated by n-1 highly variable positions. For example the LG5 motif occurrence (highlighted in red) corresponds to a pair of leucin and glycin residues which are separated by four amino acids.

sequence set. Including TMHMM predictions, each motif occurrence was assigned to a topology state as elucidated in 2.2. Additional to the defined topology states 'TM' and 'nTM', a another state has been defined for this study. Each motif, where the beginning and the end of the motif has been located in the different topology states 'TM' and 'nTM', has been assigned with the 'trans' state. Subsequent, all variable positions within each motif occurrence were examined more closely. Ultimately for each variable position the relative occurrence of each amino acid at the specified position of each motif were calculated.

To define a separation rule for the investigated motifs, an information-based approach was applied. Formally, a motif M, for instance LG5, can be interpreted as a set of variable strings with a length n. Intuitively, in case of LG5 n equals 4. To include the molecular information of the three topology states, we separated M into three motif subsets M_{TM}, M_{nTM} and M_{trans} according to the topology state S in which each single motif occurrence $m \in M$ is located. Furthermore, in each motif M_S each position pos_i with $i \in [1, n]$ can be investigated concerning its amino acid distribution. To this end, interpreting M_S as a set of strings $m_1, m_2, \ldots, m_k$ (all identified motif occurrences found in topology state S) allows formulating the relative probability $P(a|\text{pos}_i|M_S)$:

$$P(a|\text{pos}_i|M_S) = \frac{\sum_{j=1}^{k} g\left(\text{pos}_{i,m_j}, a\right)}{k} \tag{1}$$

with

$$g\left(\text{pos}_{i,m_j}, a\right) = \begin{cases} 1 & \text{pos}_{i,m_j} \text{ equals } a \\ 0 & \text{else} \end{cases}, \tag{2}$$

where a corresponds to one of the 20 canonical amino acids. To weight the significance of each probability $P(a|\text{pos}_i|M_S)$, the probability $P(a|\text{Nature})$ is applied in a log-odd formula:

$$f(a|\text{pos}_i|M_S) = \log\left(\frac{P(a|\text{pos}_i|M_S)}{P(a|\text{Nature})}\right) \tag{3}$$

The amino acid distribution $P(a|\text{Nature})$ used to test the significance of the observed relative probability at each motif position was computed from the NCBI non-redundant protein sequence set [31] (ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz).

Using these log-odd values, visualizing, clustering and information extraction can be performed. To this end, we transformed each position pos_i into a vector consisting of log-odd values which we refer to as log-odd profile $\text{LOP}\left(\text{pos}_i|M_S\right)$ and which is defined as

$$\text{LOP}\left(\text{pos}_i|M_S\right) = \begin{pmatrix} f(\text{Ala}|\text{pos}_i|M_S) \\ f(\text{Arg}|\text{pos}_i|M_S) \\ \cdots \\ f(\text{Val}|\text{pos}_i|M_S) \end{pmatrix} \tag{4}$$

Clustering all resulting $\text{LOP}\left(\text{pos}_i|M_S\right)$ was finally ensured by implementing the following distance formula:

$$\begin{aligned} &D\left(\text{LOP}\left(\text{pos}_i|M_S\right), \text{LOP}\left(\text{pos}_j|M_S\right)\right) = \\ &1 - \rho\left(\text{LOP}\left(\text{pos}_i|M_S\right), \text{LOP}\left(\text{pos}_j|M_S\right)\right) \end{aligned} \tag{5}$$

where $\rho\left(\text{LOP}\left(\text{pos}_i|M_S\right), \text{LOP}\left(\text{pos}_j|M_S\right)\right)$ corresponds to the Spearman's rank correlation coefficient. Clustering methods were applied to the LOPs to derive characteristics in motifs which determine the protein's structural and functional features.

Furthermore with these values at hand, the algorithm for predicting the topology state S based on a single motif occurrence m was implemented. At this, the pre-calculated LOPs of the corresponding motif M are employed as look-up values to compute a straight-forward winner-takes-it-all formula:

$$S = \arg \max_{\forall S \in \{\text{TM,nTM,trans}\}} \sum_{i=1}^{n} f\left(a_{m_i}|m_i|M_S\right). \tag{6}$$

The assessment of topology state prediction was performed by means of cross-validating and F-measure calculation.

By utilizing clustering methods, differences and similarities of all LOPs can be visualized and analysed in detail.

for dimensionality reduction and finally data clustering of the 20-dimensional LOP data we used the unweighted pair group method with arithmetic mean (UPGMA) [32] and the exploratory observation machine (XOM)[33]. This analysis is helpful to understand the correspondences of physicochemical properties observed in LOPs and topology states. Furthermore, this analysis enforces the found predictability of topology states. We chose the UPGMA for as visualization approach, since it is a widely used bottom-up clustering method that can be understood intuitively.

The XOM algorithm is a relatively new algorithm for dimensionality reduction. A great advantage lies in its visualization capabilities, since it can transform neighbourhood or distance relations embedded in multidimensional data into human-intelligible spaces, such as into $\mathbb{R}^2$. In literature, this property is referred to as topology-preserving mapping. However, the degree of topology-preserving mapping achieved by the XOM depends on the given problem (mainly influenced by the structure of the data and the applied distance measure) and thus the XOM output can be insufficient for analysis. In application to LOP data, however, it has shown to perform more than satisfying. Further visualizations were obtained by generating heatmaps.

3 RESULTS AND DISCUSSION

The identification of topology-discriminative positions in motifs is crucial for drawing meaningful correlations between physicochemical properties and structural and functional features. A straightforward approach to address this task is the utilization of a method to

determine the residue conservation at each variable motif position. WebLogo [34], for instance, is a widely used method to address such problems. However, WebLogo does not include any amino acid-specific background information in deriving residue conservation, since natural amino acid frequencies are not taken into account. To circumvent this problem we used LOPs for visualization instead, which, as shown in Equation 3, include natural amino acid-specific background probabilities. Essentially, this approach is quite similar as to the methods recently described in [35]. Single LOPs can be visualized as heatmaps [36] (see Figure 3) and amino acid-specific propensities at each variable position in each motif can be extracted and thus information can be gained.

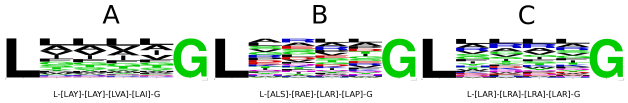


Figure 2. WebLogos [34] of the LG5 motif in the order of three topology states 'TM' (WebLogo A), 'nTM' (WebLogo B) and 'trans' (WebLogo C). However, the symbol height in each logo reflect only the relative occurrence of the corresponding amino acid. Additionally, background amino acid-specific frequencies are not taken into account which decreases the sensitivity of this method. Compared to the heatmap generated from LOPs (see Figure 3), less information can be gained. By applying WebLogo, residue propensities, with regard to the topology states, cannot be derived or identified. For instance, the leucine amino acid in 'TM' (WebLogo A) cannot be observed as more frequently at the third variable position as at other variable positions.

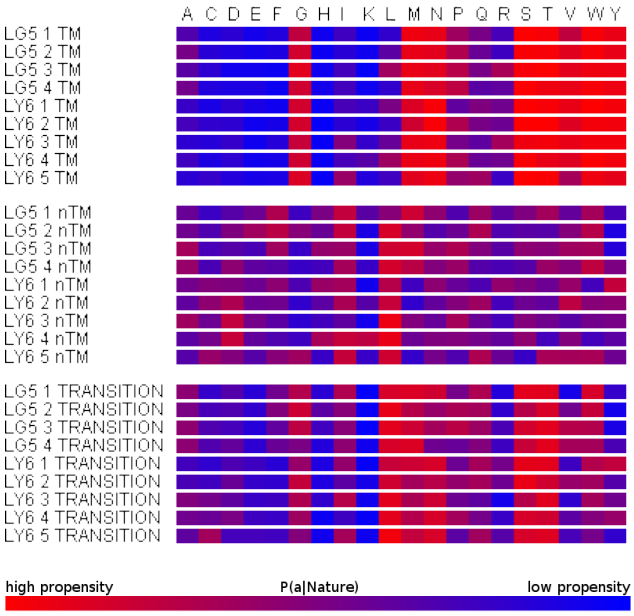


Figure 3. LOP heatmaps of the LG5 and LY6 motif. LOP heatmaps reflect the propensities of each amino acid relative to natural amino acid-specific frequencies. Increasing amino acid propensities are illustrated by an increased red color content. The below listed color scale represents the color assigning to each amino acid propensities. This visualization allows a sensitive approach to analyse amino acid propensities of each variable position of a motif according to topology states. Here, the LOP heatmap is separated by topology states, so that amino acid propensities become obvious. For example, leucin can be observed more frequently at the third variable position of transmembrane-located LG5 motifs. This results to two spatially adjacent leucin residues which form a bulky surface in transmembrane helices. Such a bulky helix surface might be important in mediating helix-helix interactions, as knob-to-hole helix packing has been reported as a key folding process in many studies (for example literature see [37; 1]).

The LOP heatmap depicted in Figure 3 shows exemplary the apparent amino acid-specific propensities according to the three topology states. Here, increasing amino acid propensities, as defined in Equation 3, are illustrated by increasing blue color content. In comparison to the WebLogos depicted in Figure 2, distinct amino acid propensities become obvious. For instance, glycine is observed more frequently in all LG5 motifs which are located in transmembrane regions. In non-transmembrane regions, the propensity of glycine is found to be reduced significantly. As a second example, in LG5 motif occurrences found in transmembrane regions, leucine is observed more frequently at the third variable position as at other variable positions. This sequence constellation results into two spatially adjacent leucine residues that form a bulky helix surface. In general, relations of topology states and the amino acid-specific propensities can be derived. This emphasizes the predictability of topology states based on single motif occurrences. The full LOP heatmap generated by this approach consists of 471 motif positions. To visualize LOP-wide correspondences, we applied UPGMA hierarchical clustering as well as the XOM algorithm. Measuring distances between LOPs was realized by utilizing Equation 5. Since 471 variable motif positions were investigated, the UPGMA-tree generated by the first approach consists of 471 leafs. To ease the analysis of the tree, the leafs were colored according to the topological state in which the corresponding motif is located. Due to the huge number of leafs, we depicted the tree only as a schematic representation which represents the observed general tree topology and identified memberships (see Figure 4). As shown, a distinct clustering, more precisely a formation of three distinct subtrees, according to the topology states is obvious. The cluster arrangement correlates to the physicochemical properties found in membrane and non-membrane located regions, since greater LOP distances are mainly dictated by the propensities of hydrophobic, hydrophilic and polar amino acids. The subtree mainly consisting of motifs located in 'trans' regions is arranged in between, which points to intermediate physicochemical motif compositions and equally distributed amino acid compositions. Similar to these findings, the XOM output (see Figure 5) shows three main clusters which correspond to the topology states too. Additionally, the cluster arrangement is found to be equal to the arrangement observed in the UPGMA-tree, where the causes of cluster formation are analogue, as well. The distinct cluster formation observed by the output of both methods point to a good separability of the variable motif positions.

A possible approach to predict the topology state of a motif from the amino acid sequence alone was implemented as elucidated in the "Information extraction and clustering" section. In this calculation, for each motif the three log-odd sums of all variable positions are computed with respect to the three topology states. The highest log-odd sum leads to the topology state winner (see Equation 6).

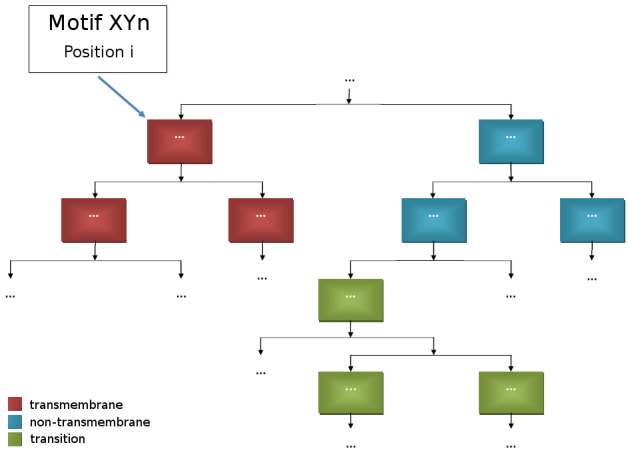


Figure 4. Schematic UPGMA-tree derived from LOP clustering: The 471 LOPs of all variable motif positions were clustered using UPGMA hierarchical clustering [32] by utilizing the LOP distance measure defined in Equation 5. Since its original size, the resulting UPGMA-tree is only depicted as a schematic. However, the tree shows three separated, distinct subtrees which correlate to the topology states the corresponding motifs are located in. The cluster arrangement corresponds to amino acid propensities and thus to physicochemical properties observed in motifs. This tree proves, that the topological location of short sequence motifs are well separable and, especially, predictable from their amino acid sequence in the variable positions.

Cross-validation was performed by excluding the evaluation set of motifs from the training motif set, which was used to generate he look-up log-odd values. In the process, each topology state winner has been assessed by F-measure. All F-measures for all investigated sequence motifs are listed in table 2. It is apparent from this table that there are motifs with high and rather small F-measures. An assessment of the resulting F-measures led to the assumption that short sequence motifs can be separated into structure-forming motifs on the one hand. Those motifs show high prediction accuracy in all investigated protein families which points to their general importance in the folding of α-helical membrane protein structures and interaction mediation of structural features. On the other hand, motifs which show high prediction accuracies only in certain families can be classified as functionally important and relevant for family-specific functional characteristics. For example, the SA3 motif observed in 'TM' in the protein family DUF2053 (PF09767) is shown a greater f-measure of $>$90% as obtained with f-measure of 46,15% in all investigated protein families. The AA3 as further motif also observed in 'TM' in the protein family DUF2243 (PF10002) is shown a greater f-measure of $>$80% as obtained f-measure of 42,58% in all investigated protein families.

4 CONCLUSION

In this work, 33 short sequence motifs reported in [18] were investigated in 32 polytopic membrane protein families with domains of unknown function. Transmembrane and non-transmembrane sequence regions were predicted using the TMHMM method [38] and topology states were annotated to all detected sequence motif occurrences. From this amino acid propensities were derived and employed to define log-odd profiles (LOP) of all variable sequence positions in the investigated motifs. Propensity tendencies according to the topology states were identified using UPGMA and XOM clustering. Both methods pointed to good separability and predictability of the topology state of a motif from its amino acid sequence. An information-based prediction algorithm was implemented and assessed using cross-validation and F-measure evaluation. Motifs showing high F-measures over all or only in certain investigated protein families were identified. From this insight, we postulate that short sequence motifs can be divided in general, structure-forming elements, which are present in numerous protein families and are highly specific to their topology location but are probably less important for functional properties. Finally, motifs showing high F-measures only in certain membrane protein families may be important elements in establishing the individual properties which are necessary for the function of the entire protein family.

Table 2. F-measures for 33 sequence motifs as assessment for topology prediction for the three different topology states 'TM', 'nTM' and 'trans'.

Motif	TM	nTM	trans
LF10	0,9900	0,9972	0,9984
LF9	0,9981	0,9966	0,9986
VF8	0,9980	0,9980	0,9957
LF8	0,9979	0,9978	0,9967
GA7	0,9972	0,9969	0,9940
PG10	0,9963	0,9996	0,9977
GY8	0,9960	0,9975	0,9953
AG7	0,9953	0,9966	0,9914
AA7	0,9946	0,9964	0,9911
LY6	0,9940	0,9947	0,9917
GG7	0,9936	0,9976	0,9888
PG9	0,9927	0,9983	0,9937
VG6	0,9893	0,9938	0,9815
SA6	0,9886	0,9963	0,9866
PG6	0,9851	0,9957	0,9808
AL6	0,9850	0,9856	0,9762
PG5	0,9538	0,9870	0,9508
GS5	0,9517	0,9772	0,9543
LG5	0,9462	0,9505	0,9140
AG5	0,9327	0,9564	0,9030
GN4	0,8673	0,9499	0,8848
IV4	0,8449	0,8423	0,7285
IL4	0,7803	0,7543	0,6342
GS4	0,7700	0,8853	0,7497
GG4	0,7652	0,8384	0,6700
SG4	0,7637	0,8788	0,7211
VL4	0,7379	0,7276	0,5842
AS4	0,7379	0,8331	0,6810
GA4	0,7305	0,7865	0,6306
AG4	0,7283	0,8013	0,6466
SA3	0,4615	0,4433	0,3831
AA3	0,4258	0,4049	0,3456
GL3	0,4086	0,3783	0,3854

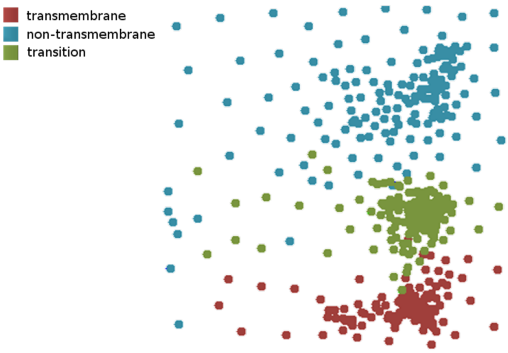


Figure 5. Output of the XOM clustering: XOM [33] is a relatively new approach for dimensionality reduction and clustering of multidimensional data. We used this approach to visualize the distance relations of the 471 investigated variable motif positions by employing the distance measure defined in Equation 5. Here, XOM delivers a two-dimensional mapping of the distance relations of all LOPs. Colored according to the topology state the corresponding motif is located in, three, well separable clusters can be seen. The LOP distances which contribute to the cluster formation are mainly dictated by the propensities of hydrophilic, hydrophobic and polar residues. Thus, the XOM output reflects physicocemical correspondences which also applies for the general cluster arrangement, with the cluster of LOPs mainly observed in 'trans' topology states (which corresponds basically to helix caps) located between the other two clusters. Similar to the UPGMA-tree depicted in Figure 4 the XOM output points to a good separability and predictability of topology states of short sequence motifs from their amino acid sequence in variable motif positions.

5 ACKNOWLEDGEMENT

The authors would like to thank the Free State of Saxony and the European Social Fond (ESF) for financial support.

REFERENCES

[1]Mary Luckey. *Membrane Structural Biology*. Cambrigde University Press, 2008.

[2]Mandy H Y. Lam and Igor Stagljar. Strategies for membrane interaction proteomics: No mass spectrometry required. *Proteomics*, 12(10):1519–1526, May 2012.

[3]D. Eisenberg, E. M. Marcotte, I. Xenarios, and T. O. Yeates. Protein function in the post-genomic era. *Nature*, 405(6788):823–826, Jun 2000.

[4]N. Lan, G. T. Montelione, and M. Gerstein. Ontologies for proteomics: towards a systematic definition of structure and function that scales to the genome level. *Curr Opin Chem Biol*, 7(1):44–54, Feb 2003.

[5]Annalisa Marsico, Dirk Labudde, Tanuj Sapra, Daniel J. Muller, and Michael Schroeder. A novel pattern recognition algorithm to classify membrane protein unfolding pathways with high-throughput single-molecule force spectroscopy. *Bioinformatics*, 23(2):e231–e236, Jan 2007.

[6]Melanie Childers, George Eckel, Alan Himmel, and Jim Caldwell. A new model of cystic fibrosis pathology: lack of transport of glutathione and its thiocyanate conjugates. *Med Hypotheses*, 68(1):101–112, 2007.

[7]Steven M. Rowe, Stacey Miller, and Eric J. Sorscher. Cystic fibrosis. *N Engl J Med*, 352(19):1992–2001, May 2005.

[8]Sandra Tan, Hwee Tong Tan, and Maxey C M. Chung. Membrane proteins and membrane proteomics. *Proteomics*, 8(19):3924–3932, Oct 2008.

[9]Glauber C. Brito and David W. Andrews. Removing bias against membrane proteins in interaction networks. *BMC Syst Biol*, 5:169, 2011.

[10]Pawel G. Sadowski, Arnoud J. Groen, Paul Dupree, and Kathryn S. Lilley. Subcellular localization of membrane proteins. *Proteomics*, 8(19):3991–4011, Oct 2008.

[11]James U. Bowie. Solving the membrane protein folding problem. *Nature*, 438(7068):581–589, Dec 2005.

[12]Gbor E. Tusnady, Zsuzsanna Dosztanyi, and Istvan Simon. Transmembrane proteins in the protein data bank: identification and classification. *Bioinformatics*, 20(17):2964–2972, Nov 2004.

[13]Gbor E. Tusnady, Zsuzsanna Dosztanyi, and Istvan Simon. Pdbtm: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res*, 33(Database issue):D275–D278, Jan 2005.

[14]A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, 305(3):567–580, Jan 2001.

[15]B. Rost, R. Casadio, P. Fariselli, and C. Sander. Transmembrane helices predicted at 95% accuracy. *Protein Sci*, 4(3):521–533, Mar 1995.

[16]David T. Jones. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*, 23(5):538–544, Mar 2007.

[17]Sikander Hayat and Arne Elofsson. Boctopus: improved topology prediction of transmembrane β-barrel proteins. *Bioinformatics*, 28(4):516–522, Feb 2012.

[18]Y. Liu, D. M. Engelman, and M. Gerstein. Genomic analysis of membrane protein families: abundance and conserved motifs. *Genome Biol.*, 3(10):research0054, Sep 2002.

[19]I. T. Arkin and A. T. Brunger. Statistical analysis of predicted transmembrane alpha-helices. *Biochim Biophys Acta*, 1429(1):113–128, Dec 1998.

[20]A. Senes, M. Gerstein, and D. M. Engelman. Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. *J. Mol. Biol.*, 296(3):921–936, Feb 2000.

[21]W. P. Russ and D. M. Engelman. The gxxxg motif: a framework for transmembrane helix-helix association. *J Mol Biol*, 296(3):911–919, Feb 2000.

[22]Alessandro Senes, Donald E. Engel, and William F. DeGrado. Folding of helical membrane proteins: the role of polar, gxxxg-like and proline motifs. *Curr Opin Struct Biol*, 14(4):465–479, Aug 2004.

[23]Roman A. Melnyk, Sanguk Kim, A Rachael Curran, Donald M. Engelman, James U. Bowie, and Charles M. Deber. The affinity of GxxxG motifs in transmembrane helix-helix interactions is modulated by long-range communication. *J Biol Chem*, 279(16):16591–16597, Apr 2004.

[24]Dirk Schneider, Carmen Finger, Alexander Prodoehl, and Thomas Volkmer. From interactions of single transmembrane helices to folding of alpha-helical membrane proteins: analyzing transmembrane helix-helix interactions in bacteria. *Curr Protein Pept Sci*, 8(1):45–61, Feb 2007.

[25]Dirk Schneider and Donald M. Engelman. Motifs of two small residues can assist but are not sufficient to mediate transmembrane helix interactions. *J Mol Biol*, 343(4):799–804, Oct 2004.

[26]Ronald Jackups, Jr and Jie Liang. Combinatorial model for sequence and spatial motif discovery in short sequence fragments: examples from beta-barrel membrane proteins. *Conf Proc IEEE Eng Med Biol Soc*, 1:3470–3473, 2006.

[27]Marco Punta, Penny C. Coggill, Ruth Y. Eberhardt, Jaina Mistry, John Tate, Chris Boursnell, Ningze Pang, Kristoffer Forslund, Goran Ceric, Jody Clements, Andreas Heger, Liisa Holm, Erik L L. Sonnhammer, Sean R. Eddy, Alex Bateman, and Robert D. Finn. The pfam protein families database. *Nucleic Acids Res*, 40(Database issue):D290–D301, Jan 2012.

[28]M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn. The Pfam protein families database. *Nucleic Acids Res.*, 40(Database issue):290–301, Jan 2012.

[29]W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, Jul 2006.

[30]Miller W. Myers E.W. Altschul S.F., Gish W. and Lipman D.J. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10, 1990. Available online at `http://toolkit.tuebingen.mpg.de/blastclust/`; visited on July 16th 2012.

[31]E. W. Sayers et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 40(Database issue):13–25, Jan 2012.

[32]Sokal R. R. and Michener C. D. A statistical method for evaluating systematic relationships. *In: University of Kansas Science Bulletin*, 38:14091438, 1958.

[33]Wismueller A. A Computational Framework for Nonlinear Dimensionality Reduction and Clustering. *Lecture Notes in Computer Science*, 5629:334–343, DOI: 10.1007/978–3–642–02397–2_38, 2009.

[34]G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner. WebLogo: a sequence logo generator. *Genome Res.*, 14(6):1188–1190, Jun 2004.

[35]M. C. Thomsen and M. Nielsen. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res.*, 40(Web Server issue):W281–287, Jul 2012.

[36]Wilkinson L. and Friendly M. The History of the Cluster Heat Map, 2008. Available online at `http://www.cs.uic.edu/~wilkinson/Publications/heatmap.pdf`; visited on July 16th 2012.

[37]D. Langosch and J. Heringa. Interaction of transmembrane helices by a knobs-into-holes packing characteristic of soluble coiled coils. *Proteins*, 31(2):150–159, May 1998.

[38]E. L. Sonnhammer, G. von Heijne, and A. Krogh. A hidden markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol*, 6:175–182, 1998.

Bibliography

[1] S. Waack and R. Merkl. *Bioinformatik Interaktiv - Algorithmen und Praxis*. WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, first edition edition, 2003.

[2] M. Zvelebil and J.O. Baum. *Understanding Bioinformatics*. Garland Science, first edition edition, 2008.

[3] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, Sep 1997.

[4] Grzegorz M. Boratyn, Alejandro A. Schaffer, Richa Agarwala, Stephen F. Altschul, David J. Lipman, and Thomas L. Madden. Domain enhanced lookup time accelerated blast. *Biol Direct*, 7(1):12, Apr 2012.

[5] D.T. Pollard and W.C Earnshaw. *Cell Biology*. Springer Verlag Berlin Heidelberg, 2007.

[6] M. Holtzhauer and J. Behlke. *Methoden der Proteinanalytik*. Springer, 1996.

[7] F. J. Burkowski. *Structural Bioinformatics. An Algorithmic Approach*. Taylor & Francis Group, LLC, first edition edition, 2009.

[8] M. Luckey. *Membrane Structural Biology - With Biochemical and Biophysical Foundation*. Cambridge University Press, 2008.

[9] E. W. Sayers et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 40(Database issue):13–25, Jan 2012.

[10] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res*, 28(1):235–242, Jan 2000.

[11] Tamotsu Noguchi and Yutaka Akiyama. Pdb-reprdb: a database of representative protein chains from the protein data bank (pdb) in 2003. *Nucleic Acids Res*, 31(1):492–493, Jan 2003.

[12] Chi Zhang, Song Liu, Hongyi Zhou, and Yaoqi Zhou. The dependence of all-atom statistical potentials on structural training database. *Biophys J*, 86(6):3349–3358, Jun 2004.

[13] M. J. Sippl. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J Comput Aided Mol Des*, 7(4):473–501, Aug 1993.

[14] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–453, Mar 1970.

[15] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197, Mar 1981.

[16] J. Ponder. TINKER – software tools for molecular design. Technical report, Dept. of Biochemistry and Molecular Biophysics, Washington University, School of Medicine, St. Louis, 2001.

[17] D. Mrozek, B. Malysiak, and S. Kozielski. EAST: Energy Alignment Search Tool. In Lipo Wang, Licheng Jiao, Guanming Shi, Xue Li, and Jing Liu, editors, *Fuzzy Systems and Knowledge Discovery*, volume 4223 of *Lecture Notes in Computer Science*, pages 696–705. Springer Berlin / Heidelberg, 2006.

[18] D. Mrozek, B. Malysiak, and S. Kozielski. An optimal alignment of proteins energy characteristics with crisp and fuzzy similarity awards. In *FUZZ-IEEE'07*, pages 1–6, 2007.

[19] D. Mrozek, B. Malysiak-Mrozek, and S. Kozielski. Alignment of protein structure energy patterns represented as sequences of fuzzy numbers. In *Fuzzy Information Processing Society, 2009. NAFIPS 2009. Annual Meeting of the North American*, 2009.

[20] James C Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D Skeel, Laxmikant Kalé, and Klaus Schulten. Scalable molecular dynamics with NAMD. *J Comput Chem*, 26(16):1781–1802, Dec 2005.

[21] F. Dressel, A. Marsico, A. Tuukkanen, M. Schroeder, and D. Labudde. Understanding of SMFS barriers by means of energy profiles. In *Proceedings of German Conference on Bioinformatics*, pages 90–99, 2007.

[22] F. Dressel. *Sequenz, Energie, Struktur - Untersuchungen zur Beziehung zwischen Primär- und Tertiärstruktur in globulären und Membran-Proteinen*. PhD thesis, Technische Universität Dresden, 2008.

[23] L. Holm and J. Park. Dalilite workbench for protein structure comparison. *Bioinformatics*, 16(6):566–567, Jun 2000.

[24] Yuzhen Ye and Adam Godzik. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, 19 Suppl 2:ii246–ii255, Oct 2003.

[25] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. Clustal w and clustal x version 2.0. *Bioinformatics*, 23(21):2947–2948, Nov 2007.

[26] Angel R. Ortiz, Charlie E M. Strauss, and Osvaldo Olmea. Mammoth (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci*, 11(11):2606–2621, Nov 2002.

[27] Alex Herbert, 2008. MaxClusterer. unpublished, Structural Bioinformatics Group, Imperial College, London. available at www.sbg.bio.ic.ac.uk/ maxcluster/index.html

[28] Z. Du, L. Li, C. F. Chen, P. S. Yu, and J. Z. Wang. G-SESAME: web tools for GO-term-based gene similarity analysis and knowledge discovery. *Nucleic Acids Res.*, 37:W345–349, 2009.

[29] M. Ashburner et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, May 2000.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, sind als solche kenntlich gemacht.

Diese Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt.

Mittweida, 18 August 2012